

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

TOWARDS ACCURATE MOLECULAR
PROPERTY PREDICTION: INNOVATIONS IN
DATA CLEANSING, MOLECULAR
DESCRIPTIONS, AND MODEL
OPTIMIZATION

TING ZHANG

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University

Department of Computing

Towards Accurate Molecular Property Prediction:
Innovations in Data Cleansing, Molecular Descriptions, and
Model Optimization

Ting ZHANG

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

September 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: _____

Name of Student: Ting ZHANG

Abstract

Predicting molecular properties with high accuracy is fundamental to accelerating progress in the discovery and optimization of functional materials. In particular, organic solar cells (OSCs) represent a rapidly advancing class of technologies for sustainable energy, characterized by low fabrication cost, mechanical adaptability, and suitability for scalable roll-to-roll manufacturing. Despite significant advancements, traditional experimental methodologies for identifying and optimizing OSC materials remain labor-intensive and time-consuming. Consequently, computational methods capable of accurately and efficiently predicting molecular properties have become increasingly essential. However, the effectiveness of these computational methods heavily depends on three interrelated aspects: the quality of the input data, the molecular descriptors utilized, and the capability of machine learning models to capture intricate molecular interactions. Recognizing this, the present thesis systematically addresses these three critical challenges through innovative strategies and novel methodologies aimed at improving predictive accuracy and efficiency in molecular property predictions, specifically targeting OSC materials.

Chapter 1 introduces the background, motivations, and objectives of this thesis, highlighting the significant potential yet persistent challenges in accurately predicting molecular and device-level properties of OSCs. It highlights the constraints of traditional computational approaches, inconsistencies in experimental data, and the limitations of current machine learning methods in effectively capturing complex molecular features and donor-acceptor interactions. To address these challenges, the thesis introduces advanced AI-driven frameworks that combine rigorous data preprocessing, novel molecular descriptors, and hierarchical transformer-based models. These innovations aim to enhance predictive accuracy, reduce dependence on extensive experimental validation, and speed up progress toward next-generation OSC materials with enhanced performance.

Chapter 2 addresses the foundational issue of input data quality, the thesis introduces a robust data cleansing framework designed to enhance the reliability of datasets used in computational modeling. In OSCs, non-fullerene acceptors (NFAs) have recently

emerged as key materials for improving device efficiency. Accurate determination of their electronic properties, such as energy levels, is crucial but remains challenging, as experimental data frequently contain inconsistencies. Traditional computational techniques, such as density functional theory (DFT), commonly exhibit prediction errors ranging from 0.2 to 0.5 electron volts (eV). To overcome these limitations, this thesis proposes a systematic data cleansing methodology. This approach involves rigorous statistical cross-validation and the exclusion of anomalous experimental entries identified based on expert knowledge. Although the reliance on expert knowledge introduces potential subjectivity and limits scalability, this process significantly reduces input errors, thereby enhancing the reliability of subsequent computational predictions. Empirical results demonstrate a substantial improvement in predictive accuracy, reducing the error margin to approximately 0.06 eV, a considerable advancement over conventional computational methods. This contribution establishes a solid foundation for accurate downstream computational modeling.

Chapter 3 develops a sophisticated molecular representation method, which constitute the core innovation of this chapter. Effective molecular descriptors are essential for accurately capturing the intrinsic properties of molecules and predicting their behavior in complex chemical environments. Existing molecular descriptors predominantly emphasize atom-level features while often neglecting higher-order structural information, such as molecular subunits or chemical motifs, that critically influence molecular properties. Although some advanced representation methods have incorporated these subunit-level characteristics, they often require extensive, labor-intensive manual labeling and feature engineering. To address these drawbacks, the thesis introduces a novel, efficient, and fully automated molecular descriptor, Ring2Vec. Inspired by advancements in natural language processing (specifically the Word2Vec mechanism), Ring2Vec conceptualizes molecular rings as the basic linguistic units, comparable to words in natural language. Through unsupervised learning, Ring2Vec is capable of deriving information from molecules across atom-level and subunit-level hierarchies, constructing a comprehensive yet computationally efficient representation of molecules. Experiments validating Ring2Vec’s performance reveal remarkable accuracy, notably achieving minimal energy-level prediction errors, significantly outperforming conventional approaches. Moreover, the generalizability of Ring2Vec suggests its potential

utility beyond OSC applications, extending to broader domains within material science and molecular property prediction tasks.

Chapter 4 optimizes machine learning models specifically tailored to molecular property prediction tasks within OSCs. While graph-based representation learning methods have gained prominence in recent years, their application in the OSC domain has been under-explored, leading to suboptimal predictive performances. OSC molecules possess distinctive structural complexities, particularly intricate ring systems, donor-acceptor interactions, and hierarchical molecular features, which conventional graph neural network (GNN) methods frequently fail to fully capture. Recognizing these limitations, the thesis develops innovative machine learning frameworks specifically designed to address the unique structural characteristics of OSC molecules. In particular, this thesis introduces RingFormer, a hierarchical transformer model on graphs, tailored to combine atomic-scale and ring-based structural information, thereby capturing the multi-scale features essential for accurate property prediction in OSCs. By fusing local message-passing modules with long-range attention mechanisms, RingFormer is capable of capturing the hierarchical complexity of molecular architectures. Comprehensive empirical evaluations conducted on multiple curated OSC datasets demonstrate RingFormer’s consistent superiority over state-of-the-art methods. Remarkably, RingFormer achieves a relative performance improvement exceeding 20% compared to competing approaches on benchmark datasets, thereby solidifying its effectiveness in accurately modeling complex molecular structures inherent in OSC materials.

Chapter 5 Building further on the concept of hierarchical modeling, the thesis culminates in the proposal of a sophisticated Hierarchical Graph Transformer with Multi-level Interactions (HGT-MI) framework, explicitly tailored to model intermolecular donor-acceptor interactions, which play a fundamental role in determining OSC device performance. Traditional computational models often overlook donor-acceptor interactions or account for only one component, leading to reduced predictive accuracy and reliability. In contrast, the HGT-MI model innovatively encodes donor-acceptor interactions through multi-scale, hierarchical representations at the atom-level, motif-level, and molecular-level. Specifically, HGT-MI leverages bi-directional multi-head cross-attention to systematically integrate donor and acceptor molecular embeddings, enabling rich and comprehensive information exchange between the two molecules. Ex-

tensive validation experiments demonstrate the substantial advantages of this approach, significantly outperforming existing models in predictive accuracy for key OSC metrics, notably power conversion efficiency (PCE). Importantly, the versatility and robustness of the HGT-MI framework highlight its broader applicability beyond OSCs. Thanks to its generalizable architecture, the framework can be applied beyond OSCs to domains like perovskite solar technologies, advanced battery chemistries, and organic light-emitting devices, underscoring its potential for broad material innovation.

Chapter 6 presents a comprehensive discussion and conclusion, synthesizing the key findings and contributions of this thesis. It traces the evolution of molecular property prediction from its roots in drug discovery to its emerging role in materials science, emphasizing the unique challenges of OSCs. The chapter integrates advances across four synergistic studies, data cleansing for noise-robust supervision, automated ring/subunit representations with Ring2Vec, hierarchical learning with RingFormer, and multi-level donor-acceptor interaction modeling with HGT-MI into a cohesive end-to-end framework for molecular and device-level prediction. Strengths and limitations are critically assessed, highlighting trade-offs between predictive accuracy, interpretability, and scalability, while also identifying open challenges such as data scarcity, morphological dynamics, and computational costs. Practical implications underscore the framework’s potential to reduce experimental dependence, accelerate NFA and donor screening, and broaden the application of AI-driven methodologies across materials domains beyond OSCs. Finally, future directions are charted, including the integration of physics-informed constraints, explicit 3D conformational modeling, and collaborative data-sharing strategies, to further enhance the scalability, robustness, and translational impact of molecular property prediction frameworks.

Publications Arising from the Thesis

1. **T. Zhang**, W. Hu, Y. Sun, C. J. Zhang, Q. Li, G. Li, and H. Yan, “Hierarchical Graph Transformer with Multi-level Interactions for Molecular Property Prediction in Organic Solar Cells”, (Under revision)
2. **T. Zhang**, K. Wang, K. Jing, G. Li, Q. Li, C. Zhang, and H. Yan, “A Ring2Vec description method enables accurate predictions of molecular properties in organic solar cells”, in *npj Computational Materials*, 2024.
3. **T. Zhang**, J. Yuk Lin Lai, M. Shi, Q. Li, C. Zhang, and H. Yan, “Data cleansing and sub-unit-based molecular description enable accurate prediction of the energy levels of non-fullerene acceptors used in organic solar cells”, in *Advanced Science*, 2024.
4. Z. Ding*, **T. Zhang*** (co-first author), Y. Li, J. Shi, and C. J. Zhang, “RingFormer: A ring-enhanced graph transformer for organic solar cell property prediction”, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 1, pp. 155–163, 2025. (Oral, top 5.64%)

Acknowledgments

Reaching the end of my PhD journey has brought with it significant challenges as well as deep rewards, and I am truly indebted to the many mentors, colleagues, friends, and institutions who guided and supported me along the way.

First and foremost, I offer my heartfelt gratitude to my supervisor, Prof. Zhang Chen, whose exceptional guidance, insightful advice, and steadfast encouragement have been pivotal throughout my doctoral research. His mentorship has not only shaped this thesis but has also inspired me to strive for excellence as a scholar. I am equally thankful to my co-supervisors, Prof. Li Qing and Prof. Li Gang, for their invaluable expertise, constructive feedback, and unwavering support. Their diverse perspectives have enriched this work and broadened my understanding of the field.

My deep appreciation also goes to my collaborator, Prof. Yan He from the Hong Kong University of Science and Technology, for his generous support and invaluable contributions; his expertise and collaborative spirit have significantly elevated the impact of this work. To my fellow lab mates and peers, thank you for the shared experiences, mutual encouragement, and countless moments of inspiration. Your companionship has made the long hours in the lab and the challenges of research far more manageable and meaningful.

My family and friends deserve special recognition for their boundless love, patience, and belief in me. To my parents, Your constant love and support have provided me with strength and stability during every stage of this journey. To my friends, thank you for sharing in my milestones and uplifting me during challenging times. Your encouragement has meant more to me than words can express.

This work is partially supported by several ongoing projects led or coordinated by Prof. Zhang Chen, including P0045948 and P0046453 (industry donations from Accel Group Holding Limited and Minshang Creative Technology Holdings Limited), P0046701 and P0046703 (PolyU internal research funding), P0048183 and P0048191 (Research Matching Grant Scheme funded by the University Grants Committee), P0048887 (Innovation and Technology Fund – ITSP, ITS/028/22FP), P0048984 (Postdoc Matching Fund Scheme), P0051906 (RGC Early Career Scheme, 25600624), and P0054482 (Two

Square Capital Limited donation).

Finally, I extend my thanks to The Hong Kong Polytechnic University for providing the resources, facilities, and opportunities that made this research possible. This thesis is a collective achievement, and I am deeply thankful to everyone who has been part of this journey.

Table of Contents

Abstract	i
Publications Arising from the Thesis	v
Acknowledgments	vi
List of Figures	xi
List of Tables	xiv
1 Introduction	1
1.1 Background and motivation	2
1.2 Overview of AI-driven molecular property prediction	5
1.2.1 Data-Driven Energy Level Prediction	7
1.2.2 Model-Driven Energy Level Prediction	9
1.2.3 Ring-Enhanced PCE Prediction	10
1.2.4 Hierarchical Graph Transformer for D-A Interaction	12
1.3 Aims and objectives	13
1.4 Structure of the thesis	16
2 Data-Driven Energy Level Prediction	19
2.1 Introduction	20
2.2 Methodologies	23
2.2.1 Data Cleansing	23
2.2.2 Subunit-Based Description	25
2.3 Model fitting and validation	30

2.4	Results	31
2.5	Discussions	38
2.6	Conclusions	40
3	Model-Driven Energy Level Prediction	42
3.1	Introductions	43
3.2	Preliminaries	46
3.2.1	Word2Vec	46
3.2.2	Graph Neural Networks	47
3.2.3	Self-attention	49
3.3	Methodologies	49
3.3.1	Ring2Vec Module	51
3.3.2	Fin2Vec Module	54
3.3.3	MG2Vec Module	55
3.4	Results	57
3.4.1	Prediction with only the Ring2Vec module.	57
3.4.2	Comparison of prediction performance of different modules or combination of modules.	62
3.5	Discussions	65
3.6	Experiments	68
3.7	Conclusions	69
4	Ring-Enhanced Graph Transformer for PCE Prediction	71
4.1	Introduction	72
4.2	Related Work	76
4.3	Problem Formulation	77

4.4	Methodologies	78
4.4.1	Hierarchical OSC Graph Construction	78
4.4.2	RingFormer Layer	81
4.4.3	Prediction Layer	85
4.5	Results	86
4.5.1	Experimental Settings	86
4.5.2	Overall Performance	91
4.5.3	Model Analysis	94
4.6	Conclusions	100
5	Hierarchical Graph Transformer with multi-level interactions	102
5.1	Introductions	103
5.2	Methodologies	106
5.2.1	Workflow of HGT-MI	106
5.2.2	Multi-level Embedding Extraction.	106
5.2.3	Multi-level Interaction Modeling.	111
5.2.4	Hierarchical Fusion and Prediction	112
5.3	Results	113
5.4	Discussion	118
5.5	Conclusions	119
6	Discussion and conclusion	120

List of Figures

2.1	Y6, a prototypical non-fullerene acceptor, is structurally composed of distinct fragments such as thiophene, pyrrole, and benzothiadiazole units, complemented by two terminal groups.	26
2.2	The calculated electronic energy levels (HOMO and LUMO) of several key aromatic motifs, typically present in non-fullerene acceptors for OSC applications, are presented.	27
2.3	Shown is the chemical configuration of a prototypical non-fused NFA structure.	28
2.4	Commonly used substituent groups attached to side chains in non-fullerene acceptors for organic solar cells.	28
2.5	Shown is the distribution of ΔHOMO and ΔLUMO . These quantities represent the differences between the HOMO/LUMO energy levels determined from CV measurements and those inferred from V_{oc} and E_{g} . Specifically: $\Delta\text{HOMO} = \text{HOMO}_{\text{CV}} - \text{HOMO}_{V_{\text{oc}}}$ and $\Delta\text{LUMO} = \text{LUMO}_{\text{CV}} - \text{LUMO}_{V_{\text{oc}}}$	32
2.6	Training results obtained from the optimized GBRT model applied to dataset #4. In (a) and (b), predicted values are plotted against experimental references for $\text{HOMO}_{V_{\text{oc}}}$ and $\text{LUMO}_{V_{\text{oc}}}$. Training points are marked as blue triangles, while test points are shown as red circles. The corresponding Pearson correlation values r are reported in the inset plots. In (c), error histograms are provided, distinguishing HOMO and LUMO.	35

2.7	Testing results obtained by applying the proposed prediction model to the “use-case” dataset. (a) Distribution of prediction errors from DFT calculations compared with experimental data; (b) distribution of prediction errors from the optimized model and curated dataset compared with experimental data; (c) comparison of predicted HOMO/LUMO values from both our model and DFT against experimental data for several representative non-fullerene acceptors.	36
3.1	Overview of the proposed framework. (a) The Ring2Vec module begins by analyzing SMILES representations to detect ring-related substructures, defined as isolated rings or fused ring pairs. These structural tokens are then transformed into numerical representations using a Word2Vec-inspired embedding procedure, generating a vocabulary of several thousand ring embeddings. (b) The MG2Vec module processes molecular graphs with a GNN to capture atom-level features. (c) The Fin2Vec module generates embeddings from molecular fingerprints. Finally, embeddings from all three modules are integrated through a self-attention mechanism to construct a unified representation of NFA molecules.	50
3.2	Depiction of the four distinct approaches used in Ring2Vec to convert molecular structures into sentence representations.	53
3.3	Experimental versus predicted HOMO and LUMO energy levels are shown using embeddings generated from (a) Fin2Vec, (b) MG2Vec, (c) Ring2Vec, and (d) their integrated representation (our proposed descriptor). In each plot, pink triangles indicate HOMO values and blue circles represent LUMO values. Predictions were obtained with the ERT model.	61
3.4	Experimental versus predicted bandgap (E_g) values derived from molecular embeddings of Fin2Vec, MG2Vec, Ring2Vec, and their integrated representation (our descriptor). Predictions were carried out with the ERT model.	62

3.5	Outlier analysis showing the chemical structures of representative NFA molecules with comparatively large prediction deviations.	66
4.1	Illustrative examples of OSC molecules	73
4.2	Overview of the RingFormer framework, shown here with two stacked RingFormer layers ($L=2$) for illustration.	79
4.3	Illustration of OSC molecules sampled from the five datasets.	87
4.4	(a) Enhancement in model performance for molecules categorized by their ring counts; (b) and (c) depict embedding maps, where color encodes the number of rings.	97
4.5	Evaluation of RingFormer’s PCE (%) predictive performance, where test MAE is reported across varying layer depths L	98
5.1	Visualization of the operational mechanism of OSC devices, showing electron-hole separation and relative energy-level alignment of donor and acceptor materials.	104
5.2	Conceptual representation of HGT-MI. (a) Multi-layer embedding derivation. (b) Cross-level interaction modeling. (c) Prediction block.	107
5.3	Effect of varied donor or acceptor motifs on device PCE.	108
5.4	Effect of removing individual components on hierarchical OSC graph performance.	116
5.5	Performance comparison between single-molecule and donor-acceptor representations.	116
5.6	PCE prediction performance of HGT-MI, shown as test MAE and MSE, for varying numbers of layers (l)	117

List of Tables

2.1	Representative cases illustrating discrepancies between energy levels obtained via DFT calculations and CV measurements.	21
2.2	Comparison of CV energy level measurements of a given material across different studies.	22
2.3	Variation of open-circuit voltage (Voc) and Y6 bandgap reported for the PM6:Y6 active layer across different studies.	24
2.4	Prediction performance of the proposed methodology, combining a subunit-based molecular representation with a GBRT model, was evaluated on four datasets of progressively refined quality. Performance was assessed using MAE, RMSE, and r , representing standard measures of error magnitude and correlation strength in AI models.	33
2.5	Evaluation of different machine learning and deep learning models, including GBRT, RF, ERF, DT, ANN, SVM, and LSTM, using SMILES strings versus subunit-based molecular descriptors as input features. Performance was evaluated in terms of MAE, RMSE and r	34
2.6	Assessment of our model relative to conventional DFT calculations, based on prediction outcomes for molecules in the chosen “use-case” dataset.	35
2.7	Summary of data for 20 representative “use-case” NFA molecules. . . .	39
3.1	Evaluation of predictive accuracy, measured by mean absolute error (MAE), across four distinct strategies for converting molecules into sentence-like representations.	57
3.2	Summary of mean absolute error (MAE) values from prediction tasks on HOMO, LUMO, and bandgap, evaluated using four molecule-to-sentence conversion methods combined with different ML models. . . .	58

3.3	Summary of prediction results, reported as mean absolute error (MAE), for the Skip-gram and CBOW variants of the Word2Vec approach. . . .	58
3.4	MAE outcomes for predicting HOMO, LUMO, and bandgap are presented, based on Word2Vec embeddings derived from Skip-gram and CBOW variants. The corresponding values are listed in the table. . . .	59
3.5	Prediction results benchmarked by MAE across molecular embeddings constructed from Fin2Vec, MG2Vec (GCN/GINE), Ring2Vec, and their combined feature space.	60
3.6	Summary of prediction performance, reported as MAE, MSE, RMSE, MAPE, R^2 , using the combined descriptor (Ring2Vec + Fin2Vec + MG2Vec) across different machine learning models.	63
4.1	Statistical overview of the dataset	88
4.2	Comparison of PCE (%) prediction performance between RingFormer and baseline methods. Results are evaluated by test MAE ($\downarrow$, where lower values indicate better performance) and reported together with the mean and standard deviation. The best outcomes are highlighted in bold , and the second best are <u>underlined</u>	91
4.3	Performance of multi-task learning on the CEPDB dataset, performance measured using test MAE ($\downarrow$; lower values reflect superior results). Units for each predicted property are listed in the second row. The best results are highlighted in bold , and the second best are <u>underlined</u>	93
4.4	Ablation results for the hierarchical OSC graph, reported in terms of test MAE ($\downarrow$, lower values indicate better performance).	94
4.5	Comparison of PCE prediction performance across alternative implementations of the ring-level graph encoder, reported in terms of test MAE ($\downarrow$, lower values indicate better performance).	95
4.6	Training and inference time per epoch (in seconds) across alternative ring-level graph encoder implementations.	95

4.7	Evaluation of performance when representing structures with either rings or BRICS-derived motifs, reported by test MAE ($\downarrow$, lower values indicate better performance).	96
4.8	Comparison of PCE prediction performance of RingFormer using alternative GNN backbones, reported in terms of test MAE ($\downarrow$, lower values indicate better performance).	98
5.1	Dataset statistics.	114
5.2	PCE prediction performance comparison of HGT-MI and baseline methods on D-A datasets. Metrics are MAE ($\downarrow$) and MSE ($\downarrow$); values are mean $\pm$ standard deviation. $\downarrow$ indicates smaller values are better.	115

Chapter 1: Introduction

The application of artificial intelligence (AI) to scientific discovery has rapidly evolved into a transformative frontier, reshaping how researchers approach complex problems in physics, chemistry, biology, and materials science [13, 11]. Within this broad paradigm of AI for Science, molecular property prediction has emerged as one of the most vibrant and impactful research areas [134, 11]. The ability to accurately predict key molecular properties, such as electronic energy levels, reactivity, optical absorption spectra, and charge transport characteristics, is pivotal for accelerating material discovery pipelines, enabling the rational design of advanced functional materials, and optimizing chemical synthesis routes [11, 143]. By providing reliable property estimates before experimental synthesis, predictive models can substantially reduce the cost, time, and environmental impact of traditional trial-and-error approaches [131, 23, 112, 163].

Historically, molecular property prediction has relied on physics-based computational methods, including quantum mechanics (QM) calculations and molecular dynamics (MD) simulations [3, 37]. These methods offer high-fidelity insights into molecular behavior and electronic structure, but they are computationally intensive [10, 44]. Density functional theory (DFT), frequently applied to determine electronic structures, for examples [56, 76, 10], becomes prohibitively expensive for large systems or for high-throughput screening of thousands of candidate molecules [44, 50, 123, 51]. Consequently, the need for faster yet reliable alternatives has become increasingly pressing, particularly in fields such as OSCs, where rapid evaluation of large molecular libraries is essential for discovering high-performance materials [50, 51].

AI-driven approaches, encompassing both machine learning (ML) and deep learning (DL), offer a compelling solution to this scalability challenge. By learning structure-property relationships directly from data, AI models can bypass the computational bottlenecks of first-principles simulations, enabling orders-of-magnitude speedups in property prediction [15, 122, 112, 129, 148]. In OSCs, these models have been successfully applied to predict crucial device parameters such as PCE, open-circuit voltage (Voc), and donor-acceptor energy-level alignment [34, 161, 86, 95].

However, despite their promise, challenges such as model generalizability, data

scarcity, and interpretability persist, hindering their full potential. This thesis aims to tackle these challenges by developing advanced AI frameworks specifically designed for robust and interpretable molecular property prediction. The proposed methodologies integrate hierarchical molecular representations, multi-level interaction modeling, and chemically-informed feature engineering to improve both accuracy and interpretability. By combining state-of-the-art deep learning architectures with domain knowledge in organic electronics, this work seeks to advance the predictive capabilities of AI models while providing mechanistic insights into structure-property relationships. Such advancements have the potential to accelerate the discovery of sustainable, high-performance materials for OSCs and other optoelectronic applications.

This chapter begins with an introduction to the research background, highlighting the transformative role of AI in molecular property prediction and its contributions to sustainable material innovation. It then reviews recent advancements in AI-driven molecular property prediction for OSCs, covering state-of-the-art model architectures, molecular feature representations, benchmark datasets, and existing limitations. Following this, the aims and objectives of this work are detailed, focusing on the development of novel AI models for robust and interpretable predictions. Finally, the research framework of this thesis is outlined, providing a structured overview of the study.

1.1 Background and motivation

The growing global demand for renewable energy highlights the critical need to develop sustainable technologies that can effectively address climate change and the ongoing energy crisis [65, 68]. Among the various renewable sources, solar energy stands out due to its cleanliness, abundance, and long-term sustainability [65]. While silicon-based photovoltaics dominate the commercial market [36, 65], emerging alternatives are being actively explored to overcome limitations in cost, flexibility, and scalability. OSCs, in particular, have attracted significant attention as a promising next-generation photovoltaic technology. Their unique advantages, including compatibility with solution processing, cost-effective large-area fabrication, and mechanical flexibility enabled by printing on lightweight substrates, make them a compelling alternative to conventional silicon-based devices [92, 154, 188]. Over the past decade, tremendous progress has been achieved, with OSC power conversion efficiencies rising from approximately 10%

to beyond 20% in state-of-the-art laboratories [121, 16, 14]. Despite these impressive advances, however, several challenges remain unresolved. Issues such as insufficient long-term stability, difficulties in large-scale manufacturing, and the slow pace of discovering and optimizing high-performance materials continue to hinder their widespread commercialization [30, 197, 5]. A major bottleneck lies in the prevailing reliance on trial-and-error molecular design and synthesis, an approach that is resource- and labor-intensive and increasingly inefficient compared with the accelerated progress in competing technologies such as perovskite solar cells [201, 120].

The performance of OSCs is inherently governed by the fine-tuned control of molecular properties. Among the most critical parameters are the energy levels, specifically, the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), and the PCE, which quantifies the fraction of sunlight that can be converted into usable electricity. These properties are strongly influenced by the underlying electronic structure of the molecules, typically defined by complex π -conjugated frameworks that include extended aromatic systems. Such architectures are crucial in facilitating charge transport and enhancing light absorption. However, the rational design of high-performance OSC materials requires not only chemical intuition but also the ability to accurately predict how subtle structural variations influence key properties. Traditional computational approaches, such as density functional theory, provide a theoretical foundation for investigating these relationships. Yet, DFT and related quantum-chemical methods are limited by high computational costs and often suffer from systematic errors due to gas-phase approximations, which render them unsuitable for large-scale screening or for capturing device-level phenomena such as PCE. Experimental methods, including cyclic voltammetry (CV), remain indispensable for characterizing energy levels, but they too present limitations: measured values can vary significantly across laboratories, with discrepancies of 0.18-0.21 eV commonly reported, thus complicating the construction of reliable models.

In light of these challenges, the application of ML and DL has proven to be a powerful strategy for advancing OSC material discovery. By leveraging the growing wealth of experimental and computational data, these approaches aim to learn structure-property relationships directly from molecular representations [11, 134, 129]. Initial efforts, such as those employing molecular fingerprints or handcrafted descriptors, demonstrated the

potential of ML to provide rapid predictions [137, 181, 34, 161]. More recent advances have introduced deep learning frameworks capable of learning molecular features in a data-driven manner, including graph neural networks (GNNs) that represent molecules as graphs of atoms and bonds [33, 43, 60, 181]. These developments have led to noticeable improvements in predictive accuracy and efficiency [161, 148]. Nevertheless, fundamental challenges remain. Many existing models rely on simplified or incomplete molecular representations, limiting their ability to capture the full complexity of OSC materials, and often require extensive domain expertise for feature engineering, which restricts scalability and generalizability [148, 181].

A particularly critical limitation is the insufficient treatment of donor-acceptor (D-A) interactions. Unlike inorganic semiconductors, OSCs rely on a heterojunction architecture where donor and acceptor molecules interact synergistically to facilitate charge separation and transport [164, 191, 171]. These interactions are central to determining photovoltaic performance. For instance, the alignment of HOMO and LUMO energy levels at the D-A interface governs exciton dissociation, while molecular compatibility and stacking arrangements influence charge mobility, open-circuit voltage, and short-circuit current density [4, 171]. Together, these factors dictate the effective energy gap and ultimately the device efficiency [145]. Current ML and DL approaches, however, predominantly treat donor and acceptor molecules as independent entities, neglecting the complex interfacial interactions that lie at the heart of OSC operation [148]. As a result, existing models fail to capture synergistic effects that arise only when donor and acceptor materials are considered jointly [148, 34]. Addressing this gap is essential for developing computational strategies that can accurately predict not only molecular-level properties but also device-level performance [148].

This thesis seeks to overcome these challenges by advancing computational approaches for OSC property prediction. The work spans multiple dimensions of the problem. First, it explores strategies for enhancing data quality, including methods for cleansing, curating, and validating OSC datasets to ensure more reliable model training and evaluation. Second, it investigates advanced molecular representations that move beyond conventional fingerprints to capture higher-order structural features, such as ring systems, conjugation patterns, and side-chain effects. Third, it develops and evaluates computational frameworks that explicitly model both single-molecule characteristics and

donor-acceptor interactions across multiple scales, from atom-level contacts to motif- and molecule-level relationships, thereby offering a more comprehensive and faithful representation of the underlying physics and chemistry. By integrating these advances, this thesis aims to bridge the gap between current modeling capabilities and the complex realities of OSC device operation, offering pathways to accelerate the discovery of high-performance materials and contribute to the broader transition toward sustainable energy technologies.

1.2 Overview of AI-driven molecular property prediction

AI-driven molecular property prediction has rapidly emerged as a transformative paradigm in contemporary materials science and chemistry, fundamentally reshaping how researchers design, evaluate, and optimize novel functional materials. By leveraging advances in ML and DL, these approaches provide a powerful alternative to traditional computational methods, which, while accurate, are often computationally prohibitive for large-scale applications. Classical methods such as DFT and MD simulations remain indispensable for providing mechanistic insights into electronic structure and dynamical behavior. However, their steep computational cost renders them impractical for high-throughput screening or large-scale exploration of chemical space. In contrast, AI-based models can learn complex, nonlinear structure-property relationships directly from data, enabling rapid and accurate predictions across diverse molecular systems. This efficiency not only accelerates the discovery pipeline but also opens up the possibility of guiding inverse design, generating candidate molecules with desired properties, thus creating new opportunities for rational materials innovation.

Several key methodological advancements have underpinned the recent progress in this field. GNNs have become particularly influential, as they represent molecules naturally as graphs, where atoms correspond to nodes and chemical bonds to edges. This graph-based abstraction allows GNNs to effectively capture both local chemical environments and global topological features, thereby modeling electronic and structural interactions with high fidelity. Models such as Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and Message Passing Neural Networks (MPNNs) have demonstrated exceptional predictive power in diverse domains, from small-molecule drug discovery to energy materials [75, 168, 43, 182]. Their flexibil-

ity in learning hierarchical molecular features has made them especially well-suited for property prediction tasks in which quantum chemical descriptors or handcrafted features may be insufficient [181, 60].

In parallel, transformer-based models have brought the power of natural language processing (NLP) to molecular representations. By treating linearized string notations of molecules (e.g., SMILES or SELFIES) as a form of chemical “language” [178, 79] models such as ChemBERTa [21], MolFormer [139], and other molecular transformers leverage transfer learning and self-supervised pretraining to capture syntactic and semantic patterns in chemical sequences [140, 57, 35, 66]. These approaches have enabled large-scale, pre-trained molecular embeddings that can be fine-tuned for specific downstream property prediction tasks, reducing the dependency on vast amounts of labeled data. Their capacity for contextualized learning has proven especially useful in predicting properties that depend on subtle chemical modifications or conformational variability.

Beyond single-modality approaches, hybrid and multimodal frameworks have emerged as powerful strategies to enhance accuracy and generalizability. By integrating complementary data sources, such as molecular graphs, textual descriptions from the literature, or experimental measurements, these models exploit synergistic information that is often overlooked by single-representation methods [84, 7, 118, 102]. For instance, combining atomistic graph features with textual descriptors extracted from scientific publications has been shown to enrich predictive performance [100, 165]. Likewise, fusing structural encodings with high-throughput spectroscopic data (e.g., diffraction and vibrational spectra) can further improve modeling fidelity and robustness [170, 212, 124, 162]. Such approaches move beyond pure structural learning and embody a more holistic view of molecular information, aligning with the inherently interdisciplinary nature of materials discovery [118].

Despite these remarkable achievements, several critical challenges continue to constrain the field. Chief among these is the scarcity of high-quality, domain-specific datasets. While large datasets exist for drug-like molecules, the availability of clean, well-annotated data for specialized classes of materials, such as organic semiconductors, catalysts, or photonic materials, remains limited. This lack of data reduces model

generalizability and often leads to overfitting when dealing with narrow material families. Moreover, the heterogeneity and inconsistency of experimental reporting across the literature further complicate model training, as discrepancies in measurement conditions, device architectures, or synthesis methods introduce uncontrolled variability into datasets.

Against this backdrop, the present thesis contributes to the evolving landscape of AI-driven molecular property prediction by focusing on a particularly critical application domain: OSCs. The progression of this thesis spans from data-driven strategies, including the curation, cleansing, and augmentation of OSC-specific datasets, to sophisticated model-driven frameworks that explicitly capture both the intrinsic characteristics of individual molecules and the complex donor-acceptor interactions at multiple hierarchical levels. Moving beyond the prediction of isolated molecular properties, this thesis focuses on device-level performance metrics, with particular emphasis on PCE. By integrating graph-based embeddings, sequence-based representations, and cross-modal interactions, it advances a comprehensive strategy for bridging the gap between molecular-scale features and device-scale outcomes. Ultimately, this thesis aims not only to improve predictive accuracy but also to establish methodological pathways that accelerate the discovery and optimization of next-generation OSC materials, thereby providing a transferable blueprint for extending AI-driven approaches to broader classes of functional materials.

1.2.1 Data-Driven Energy Level Prediction

This approach prioritizes the refinement of experimental data, leveraging data cleansing techniques to enhance the quality of inputs for predicting critical molecular properties such as the HOMO and LUMO energy levels. In the field of OSCs, accurate prediction of HOMO and LUMO levels is particularly crucial, as these frontier orbital energies directly determine the open-circuit voltage and strongly influence overall device efficiency. However, experimental measurements of these properties, most commonly obtained via cyclic voltammetry, exhibit substantial variability, with reported discrepancies ranging from 0.18 to 0.21 eV across different research groups. These inconsistencies largely stem from variations in measurement conditions, calibration protocols, and sample prepara-

tion methods. Such heterogeneity introduces significant noise into datasets and poses a major obstacle to the development of robust ML models. The challenge is further compounded in resource-limited scenarios, where datasets are relatively small (e.g., around 1000 molecules), thereby magnifying the risks of overfitting and reducing model generalizability.

To address these challenges, the data-driven methodology implements a systematic cleansing strategy in which cyclic voltammetry, derived frontier orbital energies (HOMO_{CV} and LUMO_{CV}) are cross-validated against indirectly estimated values obtained from more reliable experimental metrics, such as the open-circuit voltage (V_{oc}) and the optical bandgap. This cross-validation framework enables the identification of outliers exhibiting large discrepancies between direct CV measurements and indirect estimates. Such inconsistent data points are subsequently removed, thereby reducing noise and improving the reliability of the training dataset. Following data refinement, the curated dataset is used to construct numerical representations of non-fullerene acceptors. Each NFA is decomposed into chemically meaningful sub-units, such as conjugated aromatic rings, and flexible side chains, that play a decisive role in shaping their electronic and optoelectronic properties. By encoding these structural motifs into quantitative descriptors, the methodology captures the essential structure-property relationships in a chemically interpretable manner, while preserving the diversity of molecular design. These refined molecular representations are then employed as inputs to ML models. Among the methods evaluated, gradient boosting regression trees (GBRT) demonstrate particularly strong performance, owing to their ability to capture non-linear dependencies, handle noisy data, and maintain robustness in small-sample regimes. In comparison to linear models, GBRT provides enhanced predictive power, while being more computationally efficient and data-efficient than deep neural networks, which typically require large-scale datasets to achieve stable generalization.

The significance of this data-driven approach lies in its ability to achieve high predictive accuracy, with mean absolute errors (MAE) reaching as low as 0.06 eV for HOMO and LUMO energy levels, while simultaneously maintaining excellent computational efficiency. Importantly, predictions can be completed within seconds on standard hardware, highlighting the practicality of this method for both academic and industrial applications. Such efficiency makes it particularly valuable in resource-constrained

scenarios, where access to large-scale, high-quality datasets or advanced computational infrastructure is limited. By establishing a robust foundation of clean and reliable data, this methodology not only improves the precision of frontier orbital energy predictions but also provides a critical basis for subsequent modeling efforts that rely on high-quality inputs to achieve optimal performance. Moreover, its emphasis on systematic data refinement directly addresses one of the most pressing challenges in AI-driven materials science: the scarcity and inconsistency of domain-specific datasets. Beyond OSCs, the generalizability of this framework offers a scalable and transferable strategy, capable of being adapted to other material classes, such as catalysts, semiconductors, and photonic materials, where data quality issues similarly hinder predictive modeling. In this way, the approach contributes not only to advancing energy-level prediction in OSCs but also to shaping broader methodologies for reliable AI-driven materials discovery.

1.2.2 Model-Driven Energy Level Prediction

Advancing beyond the refinement of experimental data, this subsection introduces a model-driven paradigm that leverages pre-trained, representation-based models to capture complex molecular structures for energy level prediction. While the data-driven approach excels in enhancing dataset quality, it remains inherently constrained by the limited availability of experimental measurements and by its reliance on extensive domain expertise to account for the structural intricacies of OSC molecules. To overcome these limitations, the model-driven framework shifts the focus toward innovative molecular representations, inspired by advances in natural language processing, with the goal of improving predictive performance through scalable and automated learning strategies.

The core innovation of this approach lies in treating molecular structures as hierarchical entities, in which aromatic rings, central structural motifs in non-fullerene acceptors, are represented as fundamental units analogous to words in a sentence. This analogy enables the use of pre-training strategies on large-scale molecular databases, such as the Harvard Clean Energy Project Database (CEPDB), which contains more than two million molecules. Through unsupervised learning, the model generates expressive vector embeddings of aromatic ring systems, encoding both their topological relationships and electronic characteristics. These embeddings serve as transferable molecular

features that can be fine-tuned on smaller, experimentally derived datasets (e.g., 1300 molecules) to predict key electronic properties such as HOMO, LUMO, and bandgap (E_g) with high precision.

To further enrich predictive capacity, the pre-trained ring embeddings are integrated with complementary molecular representations, including fingerprint-based descriptors and graph-based embeddings, thereby capturing structural information across multiple length scales. For regression tasks, ensemble machine learning techniques such as extremely randomized trees are employed, providing robustness to noisy data and non-linear dependencies. This hybrid integration of unsupervised representation learning with ensemble regression balances accuracy, interpretability, and computational efficiency.

Empirical evaluations demonstrate that this model-driven approach achieves MAE of approximately 0.06 eV for energy level predictions, comparable to the best results of the data-driven method, but with enhanced robustness and transferability. By explicitly encoding aromatic ring systems as core structural motifs, the method addresses a longstanding limitation of traditional molecular fingerprints and graph neural networks, which often struggle to capture higher-order structural dependencies that are critical to OSC performance. Moreover, the adoption of unsupervised representation learning reduces dependence on expert-crafted features and detailed domain knowledge, streamlining the prediction pipeline and making it more scalable for new molecular classes. The broader impact of this framework lies in its adaptability across molecular domains. Pre-trained molecular representations can be readily transferred to other property prediction tasks, offering significant advantages in fields where high-quality experimental data remain sparse. This includes not only emerging photovoltaic materials but also broader applications in organic electronics, catalysis, and photonics. By bridging large-scale pre-training with targeted fine-tuning, this model-driven approach represents a versatile and forward-looking strategy for accelerating AI-driven molecular discovery.

1.2.3 Ring-Enhanced PCE Prediction

Extending the focus from molecular-level properties to device-level performance metrics, this subsection introduces a framework designed to enhance the prediction of PCE

in OSCs. While frontier orbital energies such as HOMO and LUMO provide critical insight into electronic structure, PCE, representing the overall efficiency of converting absorbed light into electrical energy, serves as the ultimate benchmark for device functionality. Predicting PCE, however, requires more than isolated energy-level estimation; it demands an understanding of how molecular architecture translates into collective optoelectronic behavior within complex donor-acceptor systems. In particular, the intricate arrangements of aromatic ring systems, which constitute the backbone of π -conjugated frameworks, play a pivotal role in dictating absorption spectra, exciton diffusion, and charge transport pathways. Traditional graph neural networks, though effective at capturing atom-level interactions, often fail to adequately model these higher-order structural motifs, resulting in limited accuracy for device-level predictions.

To address this challenge, the proposed method introduces a ring-enhanced hierarchical graph framework that explicitly incorporates ring-level information into molecular representations. Within this framework, molecules are encoded as multi-scale hierarchical graphs: Atom-level graphs capture local chemical bonding and atomic environments. Ring-level graphs represent aromatic systems and their interconnections, providing a mid-scale structural view. Inter-level graphs model relationships between atoms and rings, facilitating cross-scale information exchange.

A transformer-based architecture underpins this representation. Local structural information is processed through message-passing mechanisms, while multi-head cross-attention layers enable the integration of global ring-level dependencies. This dual-level encoding unifies atomistic precision with higher-order structural context, allowing the model to more effectively connect molecular-scale features with device-scale outcomes. When evaluated on large-scale datasets such as the Harvard Clean Energy Project Database, comprising approximately 2.3 million molecules, the model achieves a MAE of 0.189 eV for PCE prediction.

The key advantage of the method is its capacity to broaden AI-driven modeling from molecular descriptors toward device-scale properties, marking a substantial advance in OSC informatics. Compared to conventional GNN baselines, the ring-enhanced framework demonstrates a 22.8% improvement in predictive accuracy, underscoring the importance of explicitly modeling conjugated ring systems in OSC design. By bridging

local atomic environments with global conjugated backbones, the model captures the structural intricacies that drive device performance. Beyond PCE, the framework also establishes a foundation for predicting additional device-level parameters, including short-circuit current density (J_{sc}), fill factor (FF), and open-circuit voltage (V_{oc}), thereby offering a comprehensive platform for performance optimization. More broadly, this methodology exemplifies the versatility of graph transformers in materials science, highlighting their potential to accelerate high-throughput screening of OSC candidates, reduce dependence on costly trial-and-error experimentation, and align computational predictions with molecular design principles. Ultimately, by systematically linking structural motifs to device efficiencies, this approach provides a scalable pathway toward pushing OSC efficiencies beyond the 20% benchmark, a critical milestone for their commercial viability.

1.2.4 Hierarchical Graph Transformer for D-A Interaction

The final subsection presents the Hierarchical Graph Transformer with Multi-level Interactions, a framework explicitly designed to model the donor-acceptor intermolecular interactions that critically determine the power conversion efficiency of OSC devices. While ring-level modeling represents an important step toward capturing higher-order structural features within single molecules, it remains insufficient for device-level prediction because it overlooks the complex, hierarchical, and cross-molecular interactions that occur at the D-A interface. These include HOMO-LUMO alignment, exciton dissociation, charge transfer, and interfacial packing, all of which fundamentally govern charge separation and transport, and thereby the efficiency of OSCs.

To address this challenge, the HGT-MI framework introduces a multi-granularity representation that encodes each OSC molecule across three hierarchical levels. At the atom level, the model captures local bonding environments and fine-grained chemical features essential for describing electronic distribution and reactivity. At the motif level, it abstracts functional subunits such as aromatic rings and side chains, which govern conjugation, molecular packing, and charge transport. At the molecular level, it encodes the global structural context, integrating diverse motifs into a holistic representation of the entire molecule. This hierarchical decomposition ensures that both localized

and global structural information are explicitly preserved. Crucially, these hierarchical embeddings are not restricted to individual molecules but are integrated between donor and acceptor pairs through a bi-directional cross-attention mechanism. This design enables rich information exchange within each hierarchical level while simultaneously capturing dependencies across donor-acceptor boundaries. As a result, the framework effectively models critical interfacial processes. By allowing both fine-grained (atomic and motif-specific) and coarse-grained (molecular-level) features to interact synergistically, HGT-MI provides a comprehensive representation tailored to downstream property prediction.

Overall, HGT-MI delivers a holistic and interpretable architecture that bridges the gap between molecular-scale modeling and device-level performance prediction. Extensive benchmarking demonstrates that the framework consistently and substantially outperforms existing baselines, underscoring the importance of explicitly modeling donor-acceptor interactions. Beyond OSCs, the flexibility and scalability of HGT-MI make it broadly applicable to other material classes where intermolecular interactions are central, or instance, perovskite photovoltaic technologies and organic light-emitting diode applications, and rechargeable batteries. By advancing multi-level interaction modeling, this work establishes a methodological foundation for developing computational tools that are not only more accurate and generalizable but also scalable to increasingly complex material systems, thereby accelerating progress in AI-driven materials science.

1.3 Aims and objectives

The overarching aim of this thesis is to advance the computational prediction of both molecular- and device-level properties in OSCs through the application of artificial intelligence. OSCs represent one of the most promising photovoltaic technologies, offering advantages such as lightweight flexibility, compatibility with scalable roll-to-roll fabrication, and the potential for low-cost, sustainable energy generation. Yet, the progress of OSC research has been slowed by the limitations of conventional approaches: traditional quantum-chemical methods, while accurate, are computationally demanding and unsuitable for large-scale screening, whereas experimental discovery relies heavily on labor-intensive and time-consuming synthesis and optimization. By developing

AI-based frameworks that integrate high-quality data refinement, advanced molecular representations, and hierarchical modeling of donor-acceptor interactions, this work seeks to overcome these barriers, enabling more reliable, interpretable, and scalable predictions of material and device performance. The thesis aims not only to accelerate the discovery of high-efficiency OSC materials, but also to establish generalizable computational methodologies that reduce reliance on trial-and-error experimentation and contribute to the broader pursuit of sustainable renewable energy technologies.

The initial focus of this thesis is the improvement of data reliability in molecular property prediction, recognizing that the quality of input data fundamentally determines the accuracy and robustness of computational models. Particular emphasis is placed on data-driven strategies that refine experimental measurements, with CV-derived HOMO and LUMO energy levels serving as a primary example. By systematically cross-validating these measurements against indirect but complementary metrics, such as open-circuit voltage and optical bandgap, and filtering out data points that exhibit significant inconsistencies, this approach reduces the variability that often undermines experimental datasets. The result is a cleaner, higher-confidence dataset that serves as a more reliable foundation for machine learning models, enabling more accurate and generalizable predictions of molecular electronic properties. Importantly, this refinement process is especially valuable in resource-limited contexts, where large-scale, high-quality datasets and advanced computational resources are often unavailable, and where maximizing the utility of existing experimental data becomes critical for driving progress in materials discovery.

Building on the foundation of data refinement, this thesis investigates model-driven strategies that leverage pre-trained, representation-based approaches, drawing inspiration from advances in natural language processing. Within this framework, structural motifs such as aromatic rings and π -conjugated backbones are treated as hierarchical units, enabling the construction of expressive and chemically meaningful molecular embeddings. By first conducting large-scale pre-training on extensive molecular databases, such as the Clean Energy Project Database, and subsequently fine-tuning these models on smaller, experimentally validated datasets, it becomes possible to address the challenge of data scarcity while preserving the ability to capture the structural and electronic intricacies of OSC molecules. This paradigm demonstrates not only the potential of

transfer learning to enhance predictive accuracy in OSC research but also its scalability and adaptability across diverse molecular systems, offering a versatile path forward for AI-driven materials discovery.

The work further extends beyond molecular properties to encompass device-level characteristics, with particular focus on the prediction of power conversion efficiency. In this context, the central challenge is to establish a meaningful connection between molecular-level features and macroscopic device performance, a task that requires representations capable of capturing both atomistic details and higher-order ring-level structures that underpin charge transport and light absorption in OSCs. By integrating these multi-scale structural descriptors into computational frameworks, this direction not only enhances predictive accuracy but also provides actionable insights for the rational design of next-generation OSC materials. Ultimately, the goal is to move beyond empirical trial-and-error toward a more systematic, model-guided discovery pipeline, enabling the identification of material candidates capable of surpassing the long-sought benchmark of 20% efficiency.

Recognizing that device performance in organic solar cells is determined not only by the intrinsic properties of individual molecules but also by the complex intermolecular dynamics at play, this thesis introduces the Hierarchical Graph Transformer with Multi-level Interactions. The framework encodes each molecule at three hierarchical levels-the atom level, capturing local bonding and chemical environments; the motif level, representing functional subunits such as aromatic rings and side chains; and the molecular level, reflecting global structural context. These representations are then integrated across donor and acceptor molecules through a bi-directional cross-attention mechanism, enabling rich information exchange across molecular boundaries. By explicitly modeling processes such as electronic coupling, exciton dissociation, and charge transfer at donor-acceptor interfaces, HGT-MI provides a holistic and physically grounded representation of OSCs. This approach overcomes the inherent limitations of single-molecule models and achieves superior predictive accuracy in PCE across diverse datasets, thereby offering a scalable computational tool for device-level property prediction.

Ultimately, the thesis seeks to synthesize these approaches into a comprehensive

AI-driven framework that unifies data refinement, advanced molecular representations, and device-level modeling. The resulting system is designed to be scalable, interpretable, supporting the development of autonomous, closed-loop discovery platforms for OSCs. Beyond OSCs, the framework is envisioned to be broadly applicable to other advanced material systems, such as perovskite solar cells, organic light-emitting diodes (OLEDs), and rechargeable batteries. Through this trajectory, the thesis aspires to reduce reliance on trial-and-error methods, enhance predictive accuracy, and accelerate the development of next-generation photovoltaic and optoelectronic materials.

1.4 Structure of the thesis

This thesis is presented as a compendium of four research works conducted during the PhD study, each contributing to a comprehensive understanding of AI-driven computational methods for predicting molecular and device-level properties in organic solar cells. Each chapter builds upon the previous one, moving from data-driven refinements to model-driven innovations, and from molecular property prediction to device-property prediction. The overall structure for the main content of this thesis is designed to reflect this progression, providing a cohesive framework for advancing OSC material discovery.

Chapter 2 presents a data-driven framework for energy-level prediction, with an emphasis on improving the accuracy of HOMO and LUMO estimations for non-fullerene acceptors. The chapter lays a methodological foundation by systematically addressing data quality challenges, particularly the inconsistencies and variability inherent in cyclic voltammetry measurements. By applying machine learning techniques to curate, refine, and harmonize experimental datasets, the framework demonstrates how predictive precision can be significantly enhanced, even under conditions of limited or noisy data. This work not only underscores the importance of rigorous data preprocessing in molecular property prediction but also establishes guiding principles for constructing reliable datasets that can support scalable and reproducible advances in OSC research.

Building on this foundation, **Chapter 3** transitions to a model-driven strategy for energy level prediction, investigating pre-trained, representation-based models designed to capture the structural complexity of OSC molecules, particularly features such as aromatic ring systems and extended π -conjugation. By leveraging large-scale molecular

databases for pre-training and applying fine-tuning to smaller experimental datasets, this chapter highlights how transfer learning can effectively address data scarcity. The resulting framework provides improved accuracy and generalizability in predicting energy levels across diverse molecular datasets, offering a scalable approach to representation learning in OSC research.

Chapter 4 broadens the scope of investigation from molecular properties to device-level performance by introducing a ring-enhanced predictive model for power conversion efficiency. This framework highlights the central role of structural features, most notably ring systems and extended π -conjugation, in governing charge transport pathways, exciton generation and dissociation, and ultimately photovoltaic performance. By explicitly linking molecular-level design principles with macroscopic device behavior, the model establishes a scalable approach for more accurate PCE prediction. By this means, it not only advances predictive capability but also provides practical guidance for the rational design of OSC materials with the potential to achieve higher efficiencies.

Chapter 5 continues this progression by introducing a Hierarchical Graph Transformer with Multi-level Interactions, a novel framework specifically developed to address the intrinsic complexity of donor-acceptor interactions in OSCs. Since these interfacial interactions largely determine exciton dissociation, charge transport, and ultimately PCE, their accurate modeling is essential for bridging the gap between molecular design and device-level performance. By moving beyond the constraints of single-molecule models, HGT-MI integrates multiple scales of molecular information into a unified representation, enabling a more holistic and interpretable characterization of OSC behavior under realistic operating conditions. This design not only enhances predictive accuracy but also yields mechanistic insights into the interplay between molecular structure, morphology, and photovoltaic properties, thereby offering a valuable tool for guiding rational materials discovery.

Finally, **Chapter 6** synthesizes the findings from the preceding chapters, drawing overarching conclusions and articulating prospective directions for future research. By integrating insights from both data-driven and model-driven strategies, as well as from molecular- and device-level property prediction, this chapter proposes a unified framework for advancing OSC property prediction. It highlights the central contributions of

the thesis, situating them within the broader context of AI-driven materials discovery, while also providing a critical assessment of current limitations. By this means, it not only underscores the practical implications of the work for rational OSC design but also outlines concrete opportunities for refining computational methodologies and expanding their applicability to emerging photovoltaic systems.

Chapter 2: Data-Driven Energy Level Prediction

Summary

This chapter introduces a predictive method for estimating HOMO and LUMO levels of non-fullerene acceptors in OSCs. With an error of 0.06 eV, it significantly surpasses the accuracy of traditional DFT, whose errors typically fall between 0.2 and 0.5 eV. The approach addresses the limitations of DFT and experimental CV data variability (0.18-0.21 eV across groups) through two key components. First, data cleansing cross-validates CV-derived $\text{HOMO}_{\text{CV}}/\text{LUMO}_{\text{CV}}$ with more accurate Voc and bandgap data, using ΔLUMO ($\text{LUMO}_{\text{Voc}} - \text{LUMO}_{\text{CV}}$) to identify and remove 30% of unreliable entries from an initial dataset of 1000 NFAs, yielding a refined set of 600-700 molecules. Second, a sub-unit-based molecular description divides NFAs into aromatic sub-units (e.g., thiophene, pyrrole in Y6), encoding their electronic properties like sub-unit electronegativity (SEN) and accounting for nonfused linkages and side chains, standardized to a 30-value input for machine learning models. Among tested models, GBRT delivers the best performance (MAE 0.06 eV, RMSE 0.083 eV, r 0.994), completing predictions in seconds on a standard laptop. External validation on a "use-case" dataset of 20 commercial NFAs shows superior LUMO accuracy (MAE 0.064 eV vs. DFT's 0.347 eV) and extends to Voc prediction with a 0.06 V deviation. An online application is provided (<https://jokeviner.github.io/molecular/>), with potential for future expansion to donor polymers, making this method a valuable tool for accelerating NFA research with small datasets.

This chapter is based on "T. Zhang, J. Yuk Lin Lai, M. Shi, Q. Li, C. Zhang, and H. Yan. Data cleansing and sub-unit-based molecular description enable accurate prediction of the energy levels of non-fullerene acceptors used in organic solar cells. *Advanced Science*, 2024."

2.1 Introduction

Recent progress in organic solar cells has been primarily driven by the development of high-efficiency non-fullerene acceptors, with Y6 standing out as a representative example [194]. At present, the majority of high-efficiency OSCs (exceeding 18%) rely on Y6-derived non-fullerene acceptors [42, 177, 194, 199, 211, 25, 158]. Despite this dominant success, however, the continued development of Y6-based materials has primarily depended on empirical trial-and-error approaches, where chemical modifications are guided more by intuition and incremental experimentation than by predictive theory. This reliance has slowed the pace of molecular innovation, particularly when compared to the rapid advances observed in competing photovoltaic technologies such as perovskite solar cells [48]. The absence of robust computational methodologies capable of reliably predicting key molecular properties represents a major barrier to accelerating the rational design of next-generation NFAs, underscoring the urgent need for data- and model-driven approaches that can complement and guide experimental efforts.

Among the various material properties that determine OSC performance, accurate prediction of the frontier orbital energy levels, specifically, accurate determination of NFA frontier orbital levels (HOMO and LUMO) is crucial, given that marginal changes in these energies often lead to large variations in device efficiency [82, 195, 125, 91]. In cutting-edge PM6:Y6 systems, the HOMO level offsets are often reported to fall between 0.1 and 0.15 eV [18, 41, 198, 90, 150, 209, 29, 20]. Achieving efficiencies beyond the 20% threshold therefore requires meticulous control and minimization of donor-acceptor energy offsets, which can be realized through upward adjustments of the HOMO in steps as small as 0.05 eV or less. Such a level of precision places extraordinary demands on computational methods: the ability to forecast and fine-tune frontier orbital levels (HOMO and LUMO) is a key requirement for shaping future molecular architectures and for ensuring that synthetic work is carried out both effectively and with high impact.

In the field of OSCs, the widely adopted computational method is density functional theory calculation [55, 76, 26]. However, it has become increasingly evident that DFT calculations often suffer from significant inaccuracies when estimating the frontier orbital energies, particularly the highest occupied molecular orbital and lowest unoccupied molecular orbital, of non-fullerene acceptors [113, 169]. In prior influential

Table 2.1: Representative cases illustrating discrepancies between energy levels obtained via DFT calculations and CV measurements.

Material	LUMO _{DFT}	LUMO _{CV}	$\Delta E_V(\text{LUMO})$	HOMO _{DFT}	HOMO _{CV}	$\Delta E_V(\text{HOMO})$	
ITIC [94]	−3.41	−4.02	0.61	−5.55	−5.61	0.06	
IT-M [94]	−3.35	−3.98	0.63	−5.50	−5.58	0.08	
IT-DM [94]	−3.29	−3.93	0.64	−5.45	−5.56	0.11	
Y6 [90]	−3.52	−4.07	0.55	−5.55	−5.74	0.19	
L8-BO [90]	−3.49	−3.90	0.41	−5.53	−5.68	0.15	
L8-OD [90]	−3.49	−3.91	0.42	−5.55	−5.71	0.16	
L8-HD [90]	−3.50	−3.90	0.40	−5.54	−5.71	0.17	
DTPC-IC [186]	−3.29	−3.97	0.68	−5.03	−5.21	0.18	
DTPC-DFIC [186]	−3.44	−4.10	0.66	−5.14	−5.31	0.17	
average			0.56	average			0.14

publications, reputable groups have noted that DFT-based calculations of HOMO and LUMO energies can differ from experimental results by as much as 0.14 eV and 0.56 eV, respectively (as summarized in Table 2.1)[90, 186, 94]. Among experimental methods, CV is predominantly used in OSC studies to estimate HOMO/LUMO energies by measuring molecular redox potentials. Part of the mismatch with DFT arises because DFT addresses the gas-phase electronic configuration of individual molecules, whereas CV reflects π - π interactions in condensed solid-state assemblies [88]. (It is worth emphasizing that DFT often shows substantial inaccuracies when estimating LUMO levels, despite the fact that the acceptor’s LUMO strongly affects the open-circuit voltage V_{oc} of OSC devices). In response, a range of machine learning techniques have been applied to forecast material properties and device efficiencies in OSCs [180, 107], which have been summarized in previous review articles. Note that many past reports were focused on donor molecules used in previous generation of fullerene-based OSCs [117, 70], and there are relatively fewer reports on property prediction for state-of-the-art NFA molecules. Sun et al. proposed an approach that partitions NFA molecules into multiple fragments and applies ML algorithms to forecast OSC efficiencies rather than energy levels [203]. To our knowledge, however, no existing models have been able to predict HOMO and LUMO levels of NFAs with both accuracy and speed.

Table 2.2: Comparison of CV energy level measurements of a given material across different studies.

Material/Y6	Ref [90]	Ref [198]	Ref [29]	Ref [18]	Max–Min	SD
LUMO _(CV)	−4.07	−4.04	−4.08	−3.89	0.19	0.088
HOMO _(CV)	−5.74	−5.56	−5.70	−5.71	0.18	0.080
Material/PM6	Ref [41]	Ref [90]	Ref [150]	Ref [209]	Max–Min	SD
LUMO _(CV)	−3.64	−3.47	−3.68	−3.60	0.21	0.091
HOMO _(CV)	−5.74	−5.56	−5.70	−5.71	0.18	0.080

Another key challenge in achieving accurate predictions arises from the experimental determination of HOMO/LUMO levels, which are most commonly measured using cyclic voltammetry. While CV is widely adopted in the OSC community due to its relative simplicity and accessibility, the approach suffers from considerable variability in measurement outcomes. For instance, as illustrated in Table 2.2, CV results reported by different groups for the same material exhibit discrepancies in HOMO/LUMO values of approximately 0.18-0.21 eV [101, 149]. Such inconsistencies can propagate into downstream computational modeling, reducing both the reliability and generalizability of predictive frameworks. To mitigate this issue, it becomes essential to incorporate systematic data cleansing procedures that identify and filter out anomalous entries, thereby improving the consistency and robustness of the CV dataset. Enhancing data quality in this manner is a critical prerequisite for training machine learning models that are capable of delivering accurate and reproducible predictions of electronic properties.

In this chapter, we present a novel computational framework capable of accurately and efficiently predicting the HOMO and LUMO energy levels of non-fullerene acceptors. The first component of the approach focuses on systematic data cleansing, aimed at mitigating the experimental uncertainties inherent in cyclic voltammetry-derived HOMO/LUMO measurements (labeled as HOMO_{CV} and LUMO_{CV} in this paper). Although the HOMO_{CV}/LUMO_{CV} data can have substantial errors, other experimental parameters, including Voc and the bandgap, are generally determined with higher precision. Given their strong correlation with the HOMO/LUMO levels, these quantities can be exploited to infer the frontier orbital energies of NFAs indirectly (la-

beled as HOMO_{Voc} and LUMO_{Voc}). If the $\text{HOMO}_{\text{CV}}/\text{LUMO}_{\text{CV}}$ measurements are indeed reliable, the corresponding values should exhibit only minimal deviation from $\text{HOMO}_{\text{Voc}}/\text{LUMO}_{\text{Voc}}$. Such consistency would reinforce the validity of employing Voc-derived estimates as a trustworthy alternative to direct CV measurements. If there is a large discrepancy between $\text{HOMO}_{\text{CV}}/\text{LUMO}_{\text{CV}}$ and $\text{HOMO}_{\text{Voc}}/\text{LUMO}_{\text{Voc}}$ for a given NFA, it is likely that substantial experimental uncertainties exist, either in the $\text{HOMO}_{\text{CV}}/\text{LUMO}_{\text{CV}}$ determinations or in the Voc and bandgap values, with errors in the former being the more plausible explanation. Consequently, the deviation between $\text{HOMO}_{\text{CV}}/\text{LUMO}_{\text{CV}}$ and $\text{HOMO}_{\text{Voc}}/\text{LUMO}_{\text{Voc}}$ is adopted as a diagnostic criterion for detecting unreliable data points and enhancing dataset quality.

As the second element of our approach, we propose a molecular representation constructed at the sub-unit level, designed to capture the structural features specific to NFAs. Since NFAs typically consist of fused or linearly linked aromatic building blocks, each molecule can be systematically decomposed into fewer than ten aromatic sub-units. The electronic properties of these sub-units are then encoded as numerical descriptors, forming a compact yet chemically meaningful feature vector for input into ML models. By combining this molecular representation with the data-cleansing procedures described above, we evaluated a range of ML algorithms for predicting the HOMO and LUMO levels of NFAs. Among the tested approaches, GBRT achieved the best performance, delivering both high accuracy and computational efficiency. Specifically, the GBRT model produced predictions with a mean absolute error of approximately 0.06 eV, while completing inference within a few seconds on a standard office laptop. This level of precision and speed far surpasses conventional DFT calculations, which not only require orders of magnitude more computational resources but also typically yield errors of 0.2-0.5 eV.

2.2 Methodologies

2.2.1 Data Cleansing

To enhance the reliability of experimental data obtained from CV measurements, we implement a cross-calibration procedure that aligns CV-derived values with other experimentally reported parameters known to be more accurate. In particular, the majority

Table 2.3: Variation of open-circuit voltage (Voc) and Y6 bandgap reported for the PM6:Y6 active layer across different studies.

	Ref[90]	Ref[198]	Ref[29]	Ref[18]	Ref[41]	Ref[150]	Ref[209]	Max–Min	SD
Voc (V)	0.84	0.82	0.86	0.82	0.84	0.85	0.82	0.04	0.015
Y6 bandgap (eV)	1.33	1.35	1.33	1.33	1.33	1.32	1.35	0.03	0.010

of OSC research articles consistently report the optical bandgap E_g and Voc of solar cells. These two device parameters are generally recognized for their high accuracy and reproducibility, showing only minor discrepancies across different research groups. (Presented in Table 2.3 is a compilation of Voc and E_g data for the same PM6:Y6 material set, as reported by multiple research groups. The data demonstrate minimal variations between the different groups, with a standard deviation of only 0.010-0.015 eV for Voc and E_g as shown in Table 2.3.) It is thus clear that the Voc and E_g data have much smaller errors than HOMO/LUMO data measured by CV. Leveraging this observation, we establish a correlation between HOMO/LUMO levels and Voc/bandgap, enabling the indirect estimation of frontier orbital energies through the following relation [173]:

$$V_{oc} = HOMO_{donor} - LUMO_{acceptor} - voltage_{loss} \quad (2.1)$$

The HOMO energy levels of conventional donor materials are well documented in the literature. Although these values may exhibit noticeable variations across different research groups, it is feasible to adopt an averaged value from reported data and use this as a standard reference for each donor material. Under this assumption, the only unknown parameter in Equation 2.1 becomes the voltage loss, which is typically in the range of 0.55-0.60 eV for most high-efficiency OSCs. In this study, we adopt a fixed voltage loss of 0.60 eV. By applying Equation 2.1, the LUMO of acceptor molecules can thus be derived with relatively high precision, generally within an error margin of around 0.1 eV. Once the LUMO levels are obtained, the HOMO levels of NFAs can subsequently be determined with minimal uncertainty using the relationship below, since the experimental error in E_g is negligibly small.

$$HOMO_{acceptor} = E_g - LUMO_{acceptor} \quad (2.2)$$

To construct a reliable dataset of HOMO and LUMO energy levels, we integrated data from two complementary sources. The first source consists of values directly obtained from cyclic voltammetry measurements reported in the original literature, which we denote as HOMO_{CV} and LUMO_{CV} . The second source involves values indirectly estimated using Equation 2.1 and 2.2, based on the experimentally reported V_{oc} (open-circuit voltage) and E_{g} (bandgap) calculations, which we refer to as HOMO_{Voc} / LUMO_{Voc} in this paper. To quantitatively assess the deviations between HOMO_{CV} / LUMO_{CV} and HOMO_{Voc} / LUMO_{Voc} , we introduced two new parameters: ΔHOMO , calculated as $\text{HOMO}_{\text{Voc}} - \text{HOMO}_{\text{CV}}$, and ΔLUMO , calculated as $\text{LUMO}_{\text{Voc}} - \text{LUMO}_{\text{CV}}$. If the ΔHOMO or ΔLUMO of a certain NFA are larger than a certain threshold, we can disqualify the HOMO_{CV} / LUMO_{CV} and HOMO_{Voc} / LUMO_{Voc} and remove such problematic data from our database. By applying this quality-control and elimination strategy, we filtered out approximately 30% of the original 1000 NFAs, yielding a refined dataset of 600-700 NFAs with substantially improved reliability. This curated dataset provides a more robust foundation for subsequent machine learning modeling and predictive analysis.

2.2.2 Subunit-Based Description

NFA molecules can be generally described as a “linear combination of single or fused aromatic rings capped with two electron-withdrawing end-groups.” To facilitate a more systematic and computationally tractable description of their chemical structures, we decompose each NFA into a set of sub-units. For example, the representative Y6 molecule, illustrated in Figure 2.1, comprises seven fused aromatic rings (thiophene, pyrrole, and benzothiadiazole units) with two end groups. Then, we consider the following three aspects that can influence the HOMO/LUMO levels of the NFA molecules: the electronic properties of the individual sub-units, the nature of the linkages between these sub-units, and the effects of side chains. The details of the three aspects are described as follows.

Use Sub-Unit Electronegativity (SEN) as Numeric Inputs. Computational chemistry has established that the energy levels of NFA molecules are primarily dependent on the sub-units’ “electron negativity”. The term of electronegativity was initially introduced

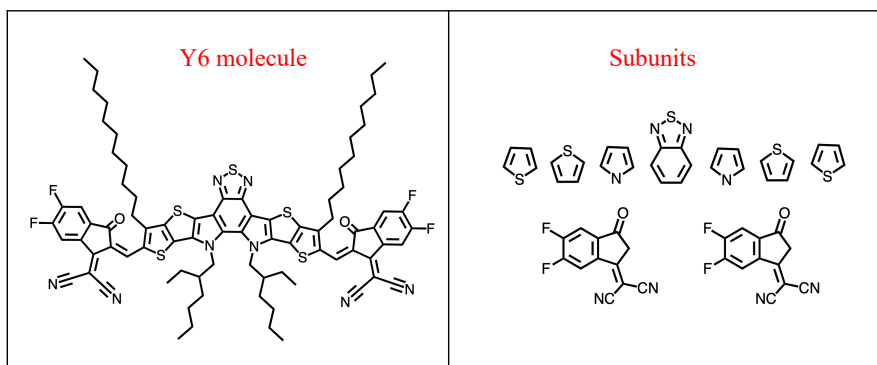


Figure 2.1: Y6, a prototypical non-fullerene acceptor, is structurally composed of distinct fragments such as thiophene, pyrrole, and benzothiadiazole units, complemented by two terminal groups.

to describe the electron-donating or electron-withdrawing effects of individual elements or atoms [64]. In this work, we extend this notion to the molecular sub-unit level by introducing the concept of sub-unit electronegativity, which captures the net donating or withdrawing tendencies of aromatic units within NFAs rather than of isolated atoms. To determine reliable initial values for SEN, we employ DFT calculations. While DFT often lacks the precision needed to predict the absolute HOMO/LUMO levels of complex NFA molecules, it remains well suited for estimating the electronic properties of smaller, structurally simpler sub-units. As depicted in Figure 2.2, we present the DFT-calculated HOMO/LUMO values for the prevalent sub-units utilized in NFAs, which will subsequently serve as the numeric descriptors for SEN, providing standardized and chemically meaningful inputs to our machine learning models.

Numeric Inputs Related to the Number of Non-Fused Linkages. The number of non-fused single-bond linkages in the molecular backbones is another crucial factor that could influence the HOMO/LUMO values of NFAs. While widely used NFAs such as Y6 and ITIC exhibit rigid, fully fused backbones with zero rotatable single bonds, a newer sub-family of non-fused NFAs has emerged, characterized by the presence of multiple rotatable linkages [12, 8, 40, 59, 93, 98]. Presented in Figure 2.3 is an illustrative instance of non-fused NFAs that contain freely rotatable single bonds. These single bonds, located in the molecular backbone and linking the four thiophene units in the central portion of the structure, significantly impact the HOMO/LUMO levels of NFAs when compared to the rigid and fully fused aromatic structure of Y6. According

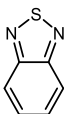
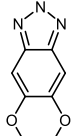
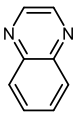
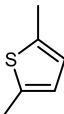
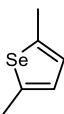
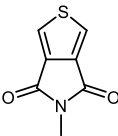
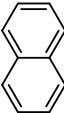
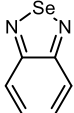
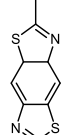
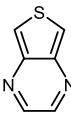
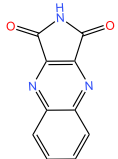
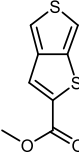
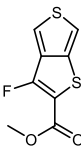
	LUMO: -0.23 HOMO: -6.35		LUMO: -0.35 HOMO: -6.32		LUMO: 0.51 HOMO: -6.12
	LUMO: -3.68 HOMO: -6.78		LUMO: 0.07 HOMO: -6.72		LUMO: -3.27 HOMO: -5.90
	LUMO: -0.38 HOMO: -5.14		LUMO: -2.35 HOMO: -6.62		LUMO: -1.88 HOMO: -7.34
	LUMO: -1.33 HOMO: -6.06		LUMO: -1.93 HOMO: -6.70		LUMO: -0.90 HOMO: -5.57
	LUMO: 3.90 HOMO: -8.32		LUMO: -0.16 HOMO: -5.76		LUMO: -2.08 HOMO: -7.09
	LUMO: -0.98 HOMO: -5.80		LUMO: -2.44 HOMO: -6.47		LUMO: -1.81 HOMO: -5.66
	LUMO: -2.28 HOMO: -6.24		LUMO: -1.43 HOMO: -6.80		LUMO: -2.75 HOMO: -7.12
	LUMO: -1.80 HOMO: -5.80		LUMO: -1.87 HOMO: -5.89		

Figure 2.2: The calculated electronic energy levels (HOMO and LUMO) of several key aromatic motifs, typically present in non-fullerene acceptors for OSC applications, are presented.

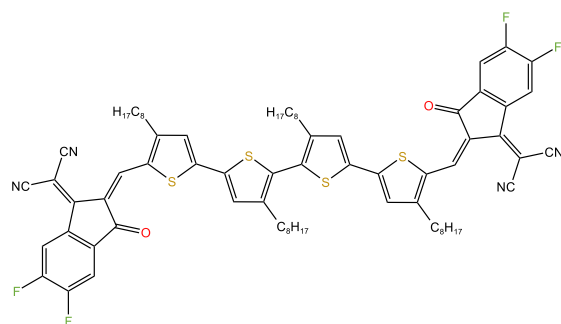


Figure 2.3: Shown is the chemical configuration of a prototypical non-fused NFA structure.

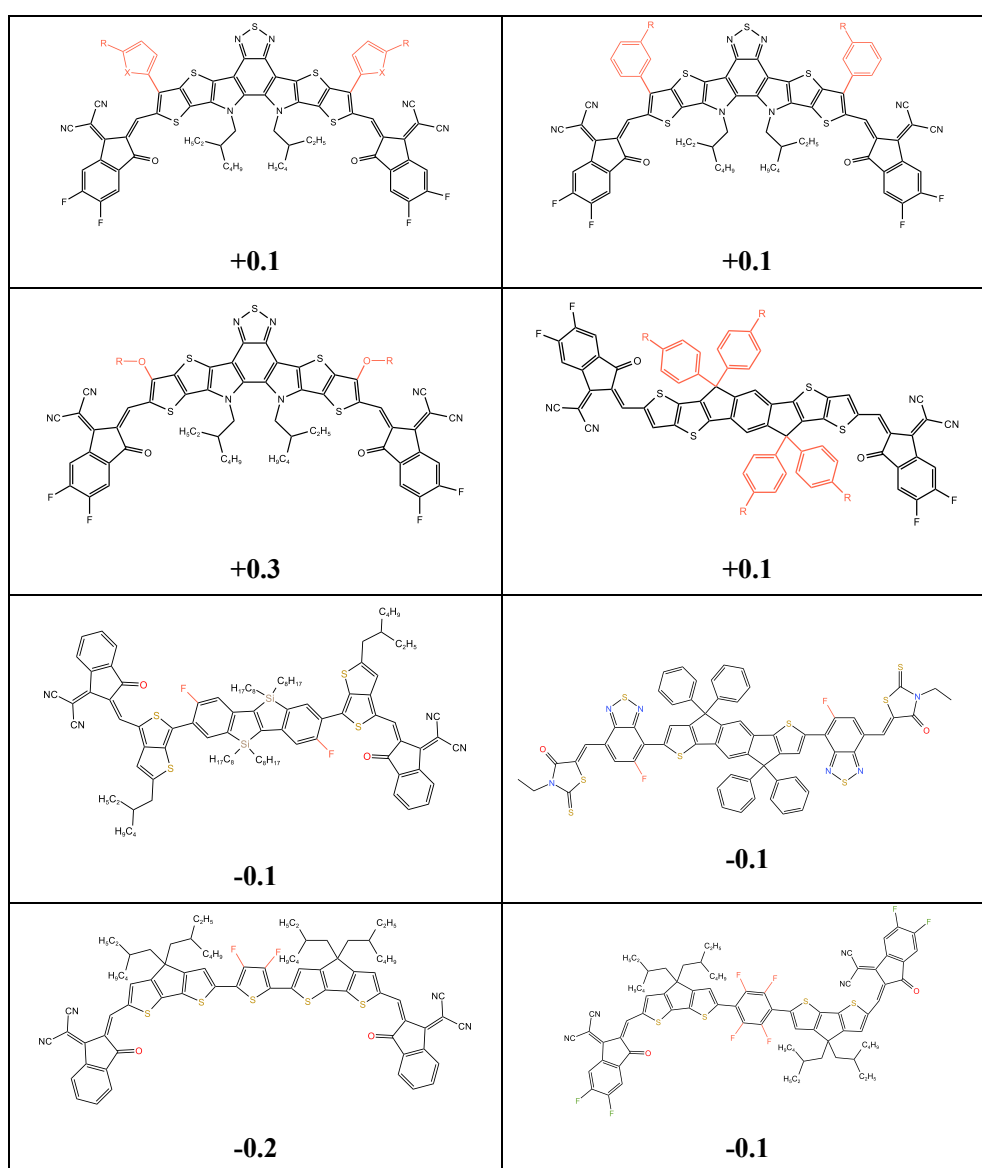


Figure 2.4: Commonly used substituent groups attached to side chains in non-fullerene acceptors for organic solar cells.

to our chemical knowledge, the presence of each non-fused single bond linkage will increase the HOMO/LUMO levels of the NFAs to some degree, which we estimate to be approximately 0.1 eV. Hence, as an input, the number of non-fused linkages and this estimated influence value (0.1 eV) should be included, in addition to the SEN values of subunits.

Numeric Inputs Related to Side Chains. Another important factor influencing the HOMO/LUMO values of NFAs is the presence of side chains or substitution groups. Chemical insights indicate that alkoxy side chains, when attached at specific positions, can significantly modulate frontier orbital energies, whereas alkyl side chains generally exert only minor effects. Building on this understanding, we estimate the contributions of various side chains and incorporate these values as input features. Table 2.4 summarizes the different side chains and substitution groups and their estimated influences in a quantitative manner, which are also used as the inputs.

Presented here is an example of how to input an NFA structure into our ML models. The commonly studied NFA molecule Y6, depicted in Figure 2.1, is composed of nine sub-units that are connected in a specific sequence, namely IC-2F, thiophene, thiophene, pyrrole, benzothiadiazole, pyrrole, thiophene, thiophene, and IC-2F. Following the calculation of the SEN values of all sub-units using DFT, we obtain a series of SEN values that represent the Y6 structure (-3.4, -0.23, -0.23, -3.68, -2.35, -3.68, -0.23, -0.23, -6.35, -6.35, -6.78, -6.62, -6.78, -6.35, -6.35, -3.4). The simplicity of these inputs for our ML models indicates that the fitting of the model should be uncomplicated.

Standardizing the Length of Numeric Inputs. Different NFAs can have different lengths of numeric inputs, which might increase the complexity of the computation. To standardize the length of the numeric inputs, we set the standard input length at 30 numeric values, which corresponds to 15 sub-units, as each unit has two numeric values for inputs. If an NFA has less than 15 sub-units, then the input of this NFA would have less than 30 numeric values. We just fill in the empty slots with zero, so that all NFAs have the same length of input. At this point, there are no NFAs reported in the OSC community, to the best of our knowledge, with more than 12 sub-units, so this standard length of 30 numeric values should cover all the possible NFAs to date. If there are possibilities to extend to more than 15 units, we can simply modify our model parameter

to accommodate the new computation needs.

2.3 Model fitting and validation

In our research, we employed a suite of seven machine learning methodologies, namely Gradient Boost Regression Trees (GBRT), Random Forests (RF), Extremely Randomized Forests (ERF), Decision Trees (DT), Artificial Neural Networks (ANN), Support Vector Machines with a Polynomial Kernel (SVM (poly)), and Long Short-Term Memory (LSTM) for predicting the energy levels of OSCs. To ensure the credibility, reliability, and robustness of our results, we adopted a K-Fold cross-validation technique as a mean to evaluate the predictive performance of these models. For the process of model training and evaluation, the entire dataset was subjected to random shuffling before being partitioned into two distinct segments - training and testing datasets. The split ratio was kept at nine to one, representative of the training to testing data, a ratio that is frequently used in the realm of machine learning. Then the prediction results of our model are evaluated by three commonly used parameters, the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), and the Pearson correlation coefficient (r). The MAE provides a straightforward measure of the average magnitude of deviations between predicted and actual values, thus reflecting the practical accuracy of the model. The RMSE, in contrast, places greater emphasis on larger errors due to the squaring operation, making it particularly useful for highlighting cases where the model fails to capture extreme values or outliers. Finally, the Pearson correlation coefficient quantifies the strength and direction of the linear relationship between predicted and true values, providing an indication of how well the model captures overall trends. The simultaneous use of these complementary metrics ensures a thorough and multidimensional evaluation of model performance, balancing considerations of accuracy, error sensitivity, and correlation with experimental values.

Similar to some of the previous studies [117], we further evaluated the performance of our model in a realistic “use-case” scenario. Specifically, we compiled a dataset of 20 widely used NFAs to compare the predictive accuracy of our approach against conventional DFT calculations. These molecules were deliberately chosen because they are commercially available, are uniquely identified by their Chemical Abstracts Service (CAS) numbers, and have both experimentally measured and DFT-computed

HOMO/LUMO energy levels reported in peer-reviewed literature. It is important to note that the experimental reference values included in this “use-case” dataset were exclusively obtained via cyclic voltammetry, which remains the most commonly employed technique in organic electronics owing to its relative simplicity, accessibility, and reliability.

2.4 Results

Among the seven AI models evaluated, the GBRT model demonstrated the best predictive performance. Consequently, we selected GBRT as the primary model for further analysis and applied it to four different datasets prepared with varying degrees of data-cleaning procedures. Table 2.4 summarizes the predictive results obtained from the GBRT model across these datasets, providing a comparative view of how different preprocessing strategies influence model performance. While dataset #1 includes $\text{HOMO}_{\text{CV}}/\text{LUMO}_{\text{CV}}$, dataset #2 includes $\text{HOMO}_{\text{Voc}}/\text{LUMO}_{\text{Voc}}$. To improve the quality of datasets #1 and #2, we use either ΔHOMO or ΔLUMO as the error indicator to identify problematic data in datasets #1 and #2. In Figure 2.5a and 2.5b, we show the distribution histograms of ΔHOMO or ΔLUMO for around 1000 NFAs, respectively. It can be seen that the distribution of ΔLUMO is slightly wider than that of ΔHOMO (standard deviations of ΔLUMO and ΔHOMO are 0.15 and 0.12 eV, respectively), which is consistent with the fact that LUMO measurement generally has larger experimental errors than HOMO. Also, the distribution of ΔLUMO is a better Gaussian distribution centered at the zero value of ΔLUMO , indicating that ΔLUMO is more influenced by random errors, while the distribution of ΔHOMO could be influenced by certain hidden systematic errors. For these reasons, we use ΔLUMO as the indicator and set the threshold at around 1 standard deviation (0.15 eV) and eliminate around 32% of data with ΔLUMO smaller than -0.15 eV or larger than 0.15 eV. Based on the ΔLUMO shown in 2.5b, dataset #3 is formed after performing this data cleansing procedure on dataset #1, while dataset #4 is formed from dataset #2.

It is worth noting that we also attempted to use ΔHOMO as the indicator for identifying problematic data. In this case, the average prediction error MAE was slightly larger than when using ΔLUMO , increasing from 0.060 to 0.064 eV. We attribute this difference to the relatively larger experimental uncertainties associated with LUMO

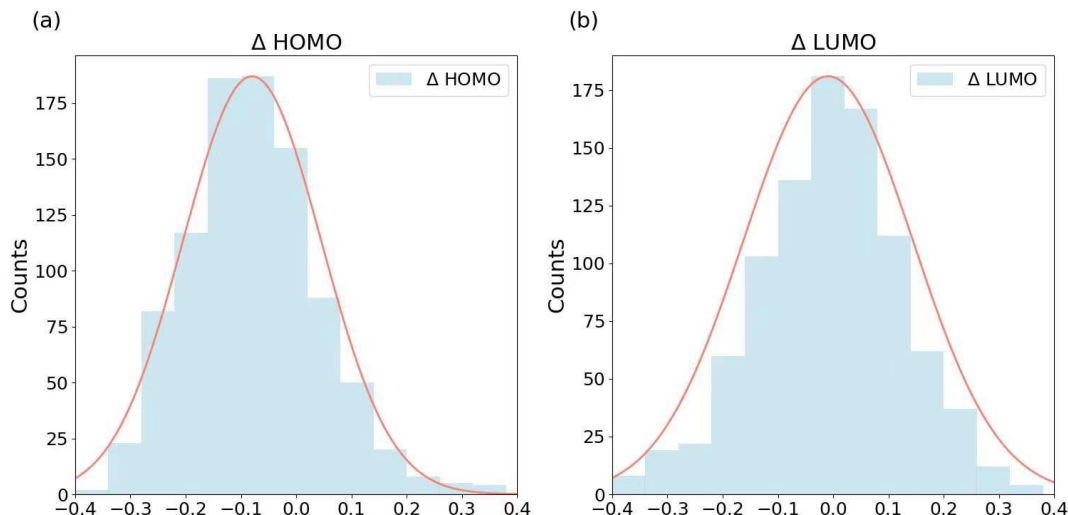


Figure 2.5: Shown is the distribution of ΔHOMO and ΔLUMO . These quantities represent the differences between the HOMO/LUMO energy levels determined from CV measurements and those inferred from V_{oc} and E_{g} . Specifically: $\Delta\text{HOMO} = \text{HOMO}_{\text{CV}} - \text{HOMO}_{V_{\text{oc}}}$ and $\Delta\text{LUMO} = \text{LUMO}_{\text{CV}} - \text{LUMO}_{V_{\text{oc}}}$.

measurements in NFA materials. As summarized in the manuscript, the typical error in HOMO values obtained from CV measurements is around 0.2 eV, whereas for LUMO values it can be as high as 0.5 eV. Consequently, using LUMO levels as the indicator provides a more effective means of identifying problematic experimental data.

The prediction results summarized in Table 2.4 include MAE, RMSE, and r , which are generally used to evaluate the performance of predictive models in a quantitative manner. These results clearly show two trends: 1) as datasets #3 and #4 outperform datasets #1 and #2, this means that our data cleansing procedure is indeed effective in reducing data error and improving the performance of prediction. In addition, the fact that datasets #2 and #4 yield better results than datasets #1 and #3 shows $\text{HOMO}_{V_{\text{oc}}}/\text{LUMO}_{V_{\text{oc}}}$ data can achieve better performance than those based on CV measurement (dataset #1). In addition, the detailed results of model training and predictions are shown in Figure 2.6. The training results show remarkably low MAE values of 0.03-0.04 eV for both HOMO and LUMO, while the test sets achieve MAE values of approximately 0.06 eV, an impressive level of accuracy compared with previous reports. As illustrated in Figure 2.6c, the majority of prediction errors fall within the range of -0.1 to 0.1 eV, and their distribution closely follows a normal Gaussian profile.

Table 2.4: Prediction performance of the proposed methodology, combining a subunit-based molecular representation with a GBRT model, was evaluated on four datasets of progressively refined quality. Performance was assessed using MAE, RMSE, and r , representing standard measures of error magnitude and correlation strength in AI models.

Computation method	Datasets	MAE [eV]	RMSE [eV]	r
Subunit description and GBRT	Dataset #1: HOMO _{CV} /LUMO _{CV}	0.082	0.116	0.990
	Dataset #2: LUMO _{Voc} , HOMO = LUMO _{Voc} + E _g	0.069	0.100	0.991
	Dataset #3: data set #1 minus 30% less accurate data	0.070	0.097	0.993
	Dataset #4: data set #2 minus 30% less accurate data	0.060	0.083	0.994

Next, we focus on the best dataset #4 to further evaluate different AI models and to compare results obtained using SMILES-based and sub-unit-based molecular representations. As shown in Table 2.5, the GBRT model achieved the best predicting results, while the other models, especially the ANN models, are not suitable for the prediction of energy levels. The enhanced performance of GBRT and RF over ANN in certain scenarios can be linked to several key factors. First, GBRT and RF, being tree-based algorithms, are intrinsically more straightforward and easier to understand than the multifaceted, nonlinear nature of ANNs. This simplicity is especially advantageous when working with datasets characterized by a limited number of features and instances; under these conditions, GBRT and RF are less likely to overfit compared to ANNs. Furthermore, GBRT and RF can seamlessly accommodate both categorical and numerical features, whereas ANNs might necessitate more involved preprocessing for these data types. Finally, the training and fine-tuning procedures for GBRT and RF are generally more time-efficient and less complicated, standing in stark contrast to the extensive computational demands and intricate tuning processes associated with ANNs.

More importantly, the results in Table 2.5 show that our sub-unit-based molecule prediction shows significantly improved prediction results compared to the SMILES-

Table 2.5: Evaluation of different machine learning and deep learning models, including GBRT, RF, ERF, DT, ANN, SVM, and LSTM, using SMILES strings versus subunit-based molecular descriptors as input features. Performance was evaluated in terms of MAE, RMSE and r .

Model	Descriptor	MAE [eV]	RMSE [eV]	r
Gradient boost regression tree (GBRT)	Sub-unit	0.060	0.083	0.994
	SMILES	0.102	0.140	0.980
Random forest (RF)	Sub-unit	0.063	0.086	0.993
	SMILES	0.100	0.138	0.981
Extremely randomized forest (ERF)	Sub-unit	0.064	0.091	0.993
	SMILES	0.099	0.138	0.981
Decision tree (DT)	Sub-unit	0.072	0.106	0.990
	SMILES	0.121	0.174	0.970
Artificial neural network (ANN)	Sub-unit	0.111	0.157	0.978
	SMILES	0.730	0.979	0.586
Support vector machine (SVM)	Sub-unit	0.089	0.112	0.991
	SMILES	0.109	0.143	0.989
Long Short-Term Memory (LSTM)	Sub-unit	0.098	0.121	–
	SMILES	0.184	0.155	0.977

based description. As the performance of models using SMILES notations as inputs can be affected by factors such as sequence sensitivity and data preprocessing. Due to the nature of SMILES notations, where a single molecule can have multiple representations, the model may face challenges in learning and recognizing a wider range of patterns, thereby reducing its accuracy and efficiency. Additionally, during data preprocessing steps like tokenization, crucial chemical information could be lost, particularly when dealing with larger molecules that exceed the model’s manageable sequence length, necessitating truncation or deletion of parts of the SMILES string. In some sense, our sub-unit-based molecular description is more suitable for the prediction of energy levels, as the electronic properties of the sub-units are the predominant factors influencing the HOMO/LUMO levels of NFAs. Therefore, the sub-unit-based description can achieve fast and accurate prediction by avoiding over-complication in computing the energy levels of large NFA molecules.

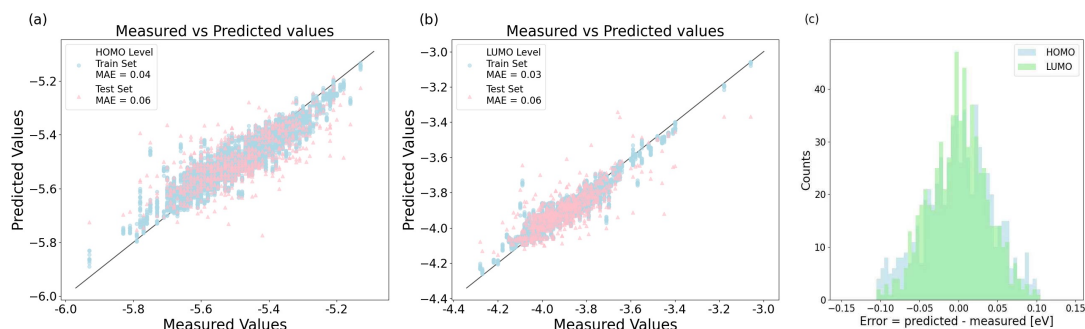


Figure 2.6: Training results obtained from the optimized GBRT model applied to dataset #4. In (a) and (b), predicted values are plotted against experimental references for HOMO_{Voc} and LUMO_{Voc} . Training points are marked as blue triangles, while test points are shown as red circles. The corresponding Pearson correlation values r are reported in the inset plots. In (c), error histograms are provided, distinguishing HOMO and LUMO.

Table 2.6: Assessment of our model relative to conventional DFT calculations, based on prediction outcomes for molecules in the chosen “use-case” dataset.

Method	MAE/RMSE of HOMO [eV]	MAE/RMSE of LUMO [eV]
This model	0.086/0.111	0.064/0.085
DFT	0.098/0.129	0.347/0.380

External Validation: Comparison with DFT Computation Using “use-case” Dataset

Inspired by the “use-case” validation approach adopted in previous studies, we further benchmarked our model’s predictive performance against conventional DFT calculations using a “use-case” dataset comprising 20 commercially available NFA molecules that are widely regarded as “workhorse” materials in the OSC research community. Experimental CV data for these NFAs are available in the literature, providing a reliable reference for evaluation. For each molecule, HOMO and LUMO levels were predicted using our ML framework and, in parallel, calculated with conventional DFT methods employing the widely adopted B3LYP hybrid functional and 6-31G advanced basis set. The results from both approaches were then directly compared against experimentally determined HOMO/LUMO values obtained from CV measurements.

Shown in Figure 2.7 is the distribution of prediction errors by DFT and our model

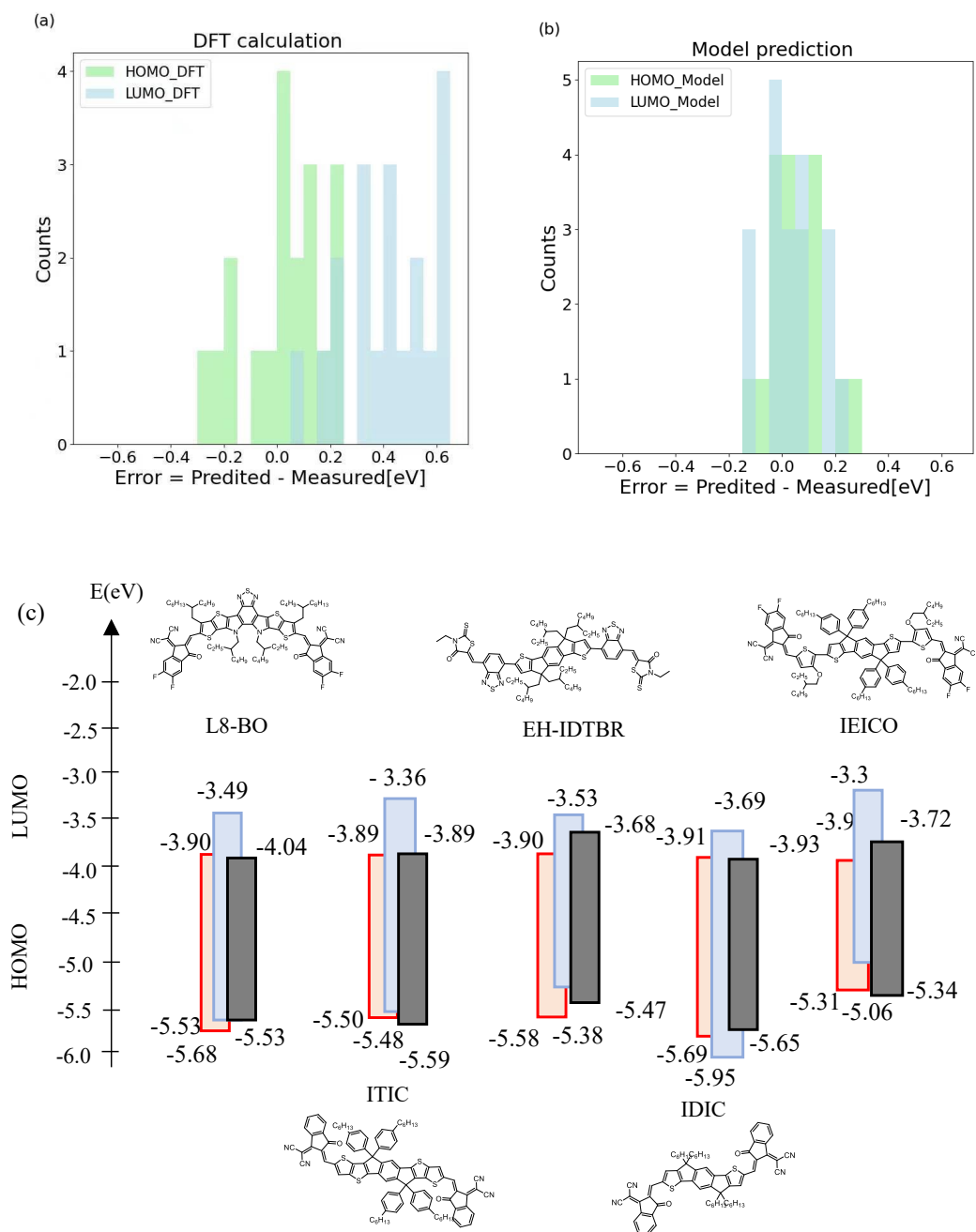


Figure 2.7: Testing results obtained by applying the proposed prediction model to the “use-case” dataset. (a) Distribution of prediction errors from DFT calculations compared with experimental data; (b) distribution of prediction errors from the optimized model and curated dataset compared with experimental data; (c) comparison of predicted HOMO/LUMO values from both our model and DFT against experimental data for several representative non-fullerene acceptors.

prediction. Our model clearly outperforms the DFT method significantly. In Table 2.6, we summarize the prediction performance of our model against conventional DFT computation. Both MAE and RMSE values are significantly reduced using our model compared to those using DFT. In specific, our model appears to be particularly effective in reducing the predictive errors for LUMO. While the MAE/RMSE of LUMO prediction is 0.347/0.380 eV using DFT computation, our method enables dramatic improvement by reducing the MAE and RMSE to 0.064 and 0.085 eV, respectively. As discussed in the introduction session, conventional DFT computation suffers from large errors for LUMO predictions. As the LUMO level of the acceptor is an important parameter that can influence the Voc of OSC devices, our model will help us to predict the Voc of OSC devices. In Figure 2.7c, we show the HOMO/LUMO levels of five commonly used NFAs by CV measurement, DFT calculations, and our model prediction. Our model prediction results appear to agree with CV measurement reasonably well for these NFAs.

From these results, it is evident that the prediction errors achieved by our method are substantially smaller than those reported in previous computational or experimental approaches. As highlighted in the introduction, earlier studies on predicting the energy levels of donor molecules in OSCs typically reported errors of around 0.2 eV. The marked reduction in error achieved by our model further validates its effectiveness. To facilitate broader accessibility, we have also developed an online application (<https://jokeviner.github.io/molecular/>) that allows researchers to utilize our prediction model. At present, this application focuses on predicting the energy levels of NFAs, reflecting the fact that most of the available data pertain to these materials. Looking ahead, we plan to extend both the model and the online platform to encompass a wider range of OSC materials, including donor polymers. We intend to distribute this tool to the OSC research community and actively gather user feedback, which will be invaluable for guiding future improvements to our prediction model.

In addition, those previous methods often rely on a large database such as the Harvard Clean Energy Project (CEP) database, which contains over 2 million molecules. Such a large database was needed because the training of the conventional AI models (using SMILES code as inputs) requires a large amount of data to achieve optimal data fitting. However, using our sub-unit-based molecular description as inputs, the model training in our study uses the data of only about 700 NFA molecules. The reason why

we can achieve accurate prediction using such a small database is probably due to the use of our sub-unit-based molecular description that simplifies molecular description and predicts the HOMO/LUMO levels more effectively. Our approach is particularly useful for NFAs, as the emergence of NFAs was quite recent, and there are only about one thousand NFAs available in the literature. In such a new research field, it is not possible to form a large database necessary for the training of conventional AI models. Therefore, our method becomes especially useful in situations where there are relatively small amounts of data available.

Lastly, our method can also be used to predict Voc. After the prediction of HOMO and LUMO levels, the predictive values can be used to estimate the Voc of the solar cells based on Equation 2.1 shown in the previous page. Certainly, the Voc of organic solar cells will also be dependent on the choices and properties of donor polymers such as HOMO. In our database, we have the average HOMO values of most donor polymers, so these values can be used to estimate the Voc of organic solar cells. We show in Table 2.7 the predicted Voc values for the 20 NFAs in the “use-case” validation. Compared with the experimental Voc values, the predicted Voc values have an average deviation of about 0.06 V, which is a respectable value comparable to the MAE of LUMO prediction. It is important to note that, in this Voc prediction, we assume that the voltage loss is 0.6 eV for all solar cell devices, which could cause some errors. If there could be a more precise prediction or estimation of voltage loss, then our Voc prediction accuracy can be further improved.

2.5 Discussions

The results presented in this chapter highlight the efficacy of our novel computational method for predicting the HOMO and LUMO energy levels of non-fullerene acceptors in OSCs, achieving a prediction error of about 0.06 eV. This represents a significant improvement over traditional density functional theory calculations, which exhibit errors of 0.2-0.5 eV, and outperforms previous machine learning approaches with errors typically around 0.2 eV. The success of our method can be attributed to the synergistic combination of data cleansing and sub-unit-based molecular description, which addresses both the variability in experimental cyclic voltammetry data and the complexity of NFA molecular structures.

The data cleansing procedure, utilizing ΔLUMO as an error indicator, effectively reduced the dataset from 1000 to 600-700 NFAs by eliminating around 32% of unreliable entries. This step is particularly critical given the larger experimental errors in LUMO measurements (up to 0.5 eV) compared to HOMO (0.2 eV), as validated by the wider distribution of ΔLUMO (standard deviation 0.15 eV) versus ΔHOMO (0.12 eV). The subsequent sub-unit-based approach, encoding electronic properties like sub-unit electronegativity (SEN) into a standardized 30-value input, simplifies the representation of complex NFA molecules, enabling the gradient boost regression trees model to achieve high accuracy with a small dataset of about 700 molecules. This is a notable advantage over conventional AI models requiring large databases, such as the Harvard Clean Energy Project with over 2 million molecules, making our method particularly suitable for the nascent field of NFA research, where data availability is limited. The

Table 2.7: Summary of data for 20 representative “use-case” NFA molecules.

#	SMA name	HOMO reported	LUMO reported	HOMO DFT	LUMO DFT	HOMO predict	LUMO predict	Donor polymer	HOMO Donor	Voc Predict	Voc Report	ΔVoc
1	BTP-eC9	-5.64	-4.05	-5.65	-3.63	-5.52	-4.06	PBDB-TF	-5.49	0.83	0.84	0.01
2	BTP-Th	-5.75	-3.96	-5.53	-3.47	-5.70	-4.00	PTQ10	-5.54	0.94	0.87	0.07
3	COi8DFIC	-5.50	-3.88	-5.66	-3.79	-5.50	-3.98	PTB7-Th	-5.22	0.64	0.68	0.04
4	EH-IDTBR	-5.58	-3.90	-5.38	-3.53	-5.47	-3.68	P3HT	-5.01	0.73	0.76	0.03
5	FBR	-5.83	-3.75	-5.88	-3.44	-5.60	-3.71	PfBT4T-2DT	-5.34	1.03	1.12	0.09
7	IDFBR	-5.75	-3.70	-5.73	-3.39	-5.49	-3.66	PfBT4T-2DT	-5.34	1.08	1.08	0.00
8	IDIC	-5.69	-3.91	-5.95	-3.69	-5.65	-3.93	PTQ10	-5.54	1.01	0.962	0.05
9	IDT-2BR	-5.52	-3.69	-5.40	-3.52	-5.44	-3.80	P3HT	-5.01	0.61	0.84	0.23
10	IEIC	-5.48	-3.85	-5.25	-3.29	-5.44	-3.79	PTB7-Th	-5.22	0.83	0.95	0.12
11	IEICO	-5.31	-3.91	-5.06	-3.30	-5.34	-3.72	J52	-5.19	0.87	0.87	0.00
12	IHIC	-5.45	-3.93	-5.66	-3.70	-5.45	-3.95	PTB7-Th	-5.22	0.67	0.754	0.09
13	INIC	-5.45	-3.88	-5.63	-3.55	-5.56	-3.82	FTAZ	-5.41	0.99	0.957	0.04
14	IT-4F	-5.66	-4.14	-5.59	-3.50	-5.61	-3.98	PBDB-T-SF	-5.38	0.80	0.88	0.08
15	IT-DM	-5.56	-3.93	-5.45	-3.29	-5.47	-3.77	PBDB-T	-5.31	0.94	0.97	0.03
16	ITIC	-5.50	-3.89	-5.48	-3.36	-5.59	-3.89	PBDB-T	-5.31	0.82	0.93	0.11
17	ITIC-Th	-5.66	-3.93	-5.61	-3.51	-5.59	-3.89	PTB7-Th	-5.22	0.73	0.80	0.07
18	IT-M	-5.58	-3.98	-5.50	-3.35	-5.59	-3.89	PBDB-T	-5.31	0.82	0.94	0.12
19	L8-BO	-5.68	-3.90	-5.53	-3.49	-5.53	-4.04	PBDB-TF	-5.49	0.85	0.87	0.02
20	Y6	-5.65	-4.10	-5.62	-3.58	-5.53	-4.04	PBDB-TF	-5.49	0.85	0.86	0.01

Average of ΔVoc : 0.06

“use-case” validation with 20 commercial NFAs further underscores the superiority of our model, reducing LUMO prediction errors from 0.347/0.380 eV (MAE/RMSE by DFT) to 0.064/0.085 eV. This improvement is crucial, as LUMO levels directly influence

Voc, and our model’s ability to predict Voc with an average deviation of about 0.06 V aligns with its LUMO accuracy. The online APP (<https://jokeviner.github.io/molecular/>) extends this utility, offering a practical tool for the OSC community, though its current limitation to NFAs suggests a need for future expansion to donor polymers.

However, several limitations warrant consideration. The assumption of a fixed voltage loss of 0.6 eV in Voc predictions may introduce errors, as actual losses vary with device architecture and materials. Additionally, the reliance on CV-derived data, even after cleansing, may inherit systematic biases from experimental conditions. The sub-unit-based description, while effective, may oversimplify intricate molecular interactions, potentially limiting generalizability to diverse OSC materials. Future work could refine voltage loss estimation, incorporate solid-state effects, and expand the dataset to enhance model robustness.

Overall, our method offers a promising framework for accelerating NFA development, particularly in data-scarce scenarios. Its integration with practical tools like the online APP positions it as a valuable asset for the OSC research community, with potential to inspire similar approaches in related fields such as perovskite solar cells.

2.6 Conclusions

To summarize, we demonstrate fast and highly accurate computation of the HOMO/LUMO levels of NFAs used in OSCs. This accomplishment was made possible through novel data cleansing procedures and machine learning techniques. Taking a data-centric approach, we first perform data cleansing on the experimental data from CV based on the correlation between problematic HOMO/LUMO data and more accurate Voc and bandgap data, which facilitated the elimination of unreliable HOMO/LUMO data and improved the data quality for machine learning model input. Next, we designed a subunit-based molecular description, which substantially accelerated computation and improved prediction accuracy. We achieved fast computation, which could be completed in just one second using a common laptop computer, while simultaneously attaining the best prediction accuracy compared to any of the previous computational methods, despite only having data on 1000 NFA molecules. As our method was successful with a smaller dataset, it has the potential to achieve even better prediction accuracy as more

molecules are developed by the OSC research community in the future. The success of our method has the potential to inspire the research community to develop similar computational models to predict other important molecular properties, such as solubility, charge mobility, and device efficiency of OSCs.

Chapter 3: Model-Driven Energy Level Prediction

This chapter introduces Ring2Vec, a novel molecular description method for predicting the properties of non-fullerene acceptors in OSCs. Addressing the limitations of traditional computational methods like DFT, which are computationally intensive and error-prone, and prior machine learning approaches requiring labor-intensive data labeling, Ring2Vec leverages both atom-level and subunit-level molecular features. Inspired by the Word2Vec model from natural language processing, Ring2Vec treats molecular rings as “words” and molecules as “sentences,” enabling automatic feature extraction from SMILES strings without manual labeling. The method achieves a prediction error of about 0.06 eV for HOMO/LUMO levels, outperforming conventional methods (errors of 0.2-0.5 eV). The methodology integrates three modules: Ring2Vec, Fin2Vec, and MG2Vec. Ring2Vec uses unsupervised pre-training on the Harvard CEP database to convert ring-based molecular features into high-dimensional vectors, employing four molecule-to-sentence conversion approaches. Fin2Vec processes SMILES strings into molecular fingerprints using the RDKit library, while MG2Vec employs GNNs to capture atom-level interactions. The combined approach achieves the lowest errors when paired with the Extremely Randomized Trees (ERT) model, surpassing standalone GNN models like GCN and GINE. The approach excels with limited data, making it ideal for the nascent NFA field, and its automated feature extraction enhances scalability.

This chapter is based on ”T. Zhang, K. Wang, K. Jing, G. Li, Q. Li, C. Zhang, and H. Yan, ‘A Ring2Vec description method enables accurate predictions of molecular properties in organic solar cells’, in *npj Computational Materials*, 2024.”

3.1 Introductions

Organic solar cells have attracted considerable attention as a promising photovoltaic technology, owing to their lightweight nature, mechanical flexibility, and low-cost fabrication. Unlike conventional inorganic solar cells, OSCs can be processed from solution and manufactured using scalable techniques such as roll-to-roll printing, enabling the production of flexible and large-area devices. These unique advantages position OSCs as strong candidates for next-generation solar energy applications, particularly in portable, wearable, and building-integrated photovoltaics [136, 71, 90, 99, 61]. However, the advancement of OSCs has been comparatively slower than that of competing solar technologies, such as perovskite solar cells, largely due to the trial-and-error nature of molecular design and synthesis. This labor-intensive approach not only hinders the pace of discovering high-performance materials but also limits the systematic understanding needed to achieve consistent improvements in device efficiency [193, 194, 42, 158, 177, 199, 211]. For instance, non-fullerene acceptors, which serve as crucial components of OSCs, are typically obtained through complicated multi-step synthetic routes. Designing such molecules is further complicated by the difficulty of accurately predicting their optoelectronic properties and structure-property relationships, making the discovery of new high-performance NFAs both time-consuming and resource-intensive [103, 70, 141, 117]. Consequently, there is an urgent need to develop computational models capable of rapid and precise predictions of the properties of NFA molecules. Traditional computational chemistry techniques, such as DFT calculations [55, 76, 26], often yield large prediction errors and require substantial computational resources [113, 169]. Recent advancements in artificial intelligence and machine learning present an opportunity to surpass the capabilities of conventional computational chemistry methods by developing efficient ML models for NFA property predictions [159, 127, 174].

To predict the properties of organic molecules, including NFAs, the first step involves converting molecular structures into numerical representations that can be processed by computer algorithms. Among the most widely adopted approaches is the Simplified Molecular Input Line Entry System (SMILES), which encodes molecular structures as linear ASCII strings. This representation allows molecules to be handled

in a machine-readable format and has been extensively used in cheminformatics and machine learning tasks for property prediction and virtual screening [178]. However, SMILES is essentially an atom-level description method that, while effective for representing connectivity, often overlooks higher-level molecular information such as subunit features and structural motifs. To address this limitation and capture more chemically meaningful patterns, one common strategy is to convert SMILES strings into molecular fingerprints. These fingerprints provide fixed-length numerical vectors that encode the presence or absence of particular substructures or functional groups, thereby enabling machine learning models to incorporate higher-level molecular information beyond simple atom-to-atom connectivity [127, 126, 115], which are specific molecular features that can have significant impacts on molecular properties. In another attempt, researchers treat molecules as Molecular Graphs using Graph Neural Networks to predict their properties [110, 138, 60, 43, 156, 6]. However, both the fingerprint and MG approaches are not effective in representing high-level molecular features such as subunit or aromatic rings, which are important components of NFA molecules [60, 34].

To overcome the limitations of atom-level and fingerprint-based methods, another promising approach to molecular description has emerged, in which researchers represent NFA molecules at the subunit level rather than at the level of individual atoms. In this framework, expert knowledge from chemistry is leveraged to manually identify, designate, and label the functional subunits that are commonly found within NFA structures. These subunits often corresponding to chemically meaningful building blocks such as donor cores, acceptor end groups, or π -bridges, are then used collectively to describe the entire molecule in a more interpretable manner. For example, in the study by Sun and co-workers, the authors employed chemical intuition to isolate key functional subunits within NFAs and encoded these units as one-hot vectors, thereby creating a simplified yet chemically meaningful representation of the molecular architecture [203]. In the previous chapter, we use DFT-calculated energy levels to label the properties of subunits and then use these subunit properties to describe the whole molecule [204]. By incorporating chemically meaningful subunit information, such approaches have been shown to markedly enhance prediction accuracy compared to purely atom-level or string-based representations. Nevertheless, a major drawback lies in their reliance on manual identification and labeling of subunits, which depends heavily on expert

chemical knowledge. This process is not only highly labor-intensive but also difficult to scale, thereby restricting the broader applicability of such methods to larger datasets or more diverse classes of OSC molecules.

In this chapter, we introduce a novel molecular description method designed to address the key limitations of existing approaches. Unlike previous methods that relied exclusively on atom-level descriptors or labor-intensive subunit labeling, our framework seamlessly integrates both atom-level and subunit-level information into a unified representation. By doing so, it not only captures fine-grained structural details but also encodes higher-level chemical motifs that are critical for accurately predicting molecular properties. An additional advantage of our approach lies in its efficiency: molecular features are automatically extracted directly from readily available SMILES strings, completely eliminating the need for manual subunit identification and expert-driven labeling. As a result, our method provides a more comprehensive and chemically meaningful description of OSC molecules, while offering exceptional scalability, robustness, and broad applicability across diverse molecular datasets.

The key innovation of our method lies in a novel Ring2Vec description method, drawing inspiration from the widely utilized Word2Vec algorithms in natural language processing [213, 114]. As aromatic rings are the most important subunits of NFA materials, our Ring2Vec model is designed to effectively represent the information on the “rings” in organic molecules. Similar to the effectiveness of the Word2Vec model in representing the true meanings of words within their contexts, our Ring2Vec interprets the “rings” within NFA molecules as “words” and considers the entire molecule as a “sentence”. This innovative approach enables us to effectively represent the information on aromatic rings in NFA molecules. Importantly, by integrating our Ring2Vec module with the aforementioned fingerprint and molecular graph methodologies (referred to as Fin2Vec and MG2Vec in the following), we establish a molecular description framework that offers a comprehensive representation of diverse molecular features across atom, subunit, and molecular levels. Leveraging this integrated framework, we achieve fast and accurate predictions of the energy levels of NFA molecules, with a minimal prediction error of merely 0.06 eV, markedly surpassing the performance of conventional computational chemistry methods [117].

Furthermore, our method can automatically extract useful molecular features from the commonly available SMILES strings and thus should have broad applicability across various domains of molecular description and property prediction.

3.2 Preliminaries

In this section, we first introduce the Word2Vec methodology, which provides the conceptual foundation for our proposed Ring2Vec method. The principles underlying Word2Vec, particularly its ability to capture contextual relationships within sequential data, serve as an important inspiration for extending similar ideas to molecular substructures. Furthermore, to situate our work within the broader field of molecular machine learning, we also present a brief overview of GNNs for molecular representation. This background discussion highlights the relevance of GNN-based approaches to our study and underscores their role in shaping recent advances in molecular property prediction.

3.2.1 Word2Vec

Word2Vec is a widely used algorithm in natural language processing that generates semantic vector representations of words. These vector embeddings capture not only the intrinsic meaning of words but also their contextual relationships, thereby enabling a more nuanced and computationally tractable understanding of language. The algorithm achieves this by analyzing word co-occurrence patterns within large-scale text corpora, employing predictive models to infer embeddings from the surrounding linguistic context. Once trained, Word2Vec provides a versatile representation that has been successfully applied to a wide range of downstream NLP tasks, including sentiment analysis, named entity recognition, machine translation, and text summarization. Within the Word2Vec framework, two primary modeling approaches are typically employed to construct word embeddings:

1. **Continuous Bag of Words (CBOW):** In this approach, the model predicts a target word based on its surrounding context words. The order of the context words is not taken into account, making it conceptually similar to the traditional bag-of-words model. Despite this simplification, CBOW is computationally efficient and well-suited for large datasets.

2. **Skip-gram:** In contrast, the Skip-gram approach reverses the prediction task by using a given target word to predict its surrounding context words. This method places greater emphasis on words that are closer in proximity to the target, enabling it to capture fine-grained semantic and syntactic relationships more effectively.

By employing these two modeling approaches and pre-training on a large-scale text corpus, Word2Vec is able to capture rich linguistic patterns, encoding word relationships as meaningful vector offsets in the embedding space. As a result, the learned vector representations effectively preserve both syntactic and semantic relationships among words, thereby providing a powerful foundation for a wide range of downstream natural language processing tasks, exemplified by $\text{vec}(\text{"China"}) - \text{vec}(\text{"Beijing"}) \approx \text{vec}(\text{"Japan"}) - \text{vec}(\text{"Tokyo"})$, and the compositionality of words, such as $\text{vec}(\text{"Germany"}) + \text{vec}(\text{"capital"})$ closely resembling $\text{vec}(\text{"Berlin"})$. Therefore, Word2Vec simultaneously educates on various aspects of language, encompassing word meanings and their grammatical aggregations.

In our study, we draw an analogy between molecules and sentences, where ring-based subunits are treated as the fundamental "words" of the molecular language. By applying the Word2Vec framework, our aim is to learn dense, real-valued vector representations of these subunits, capturing both their intrinsic characteristics and their contextual relationships within the molecular structure. This representation not only enriches our understanding of molecular rings but also provides a more expressive and chemically meaningful depiction of their structural and functional attributes.

3.2.2 Graph Neural Networks

Graph Neural Networks leverage the inherent graph structure and incorporate the features of nodes and edges to generate embeddings for individual nodes, denoted as v , within a given graph G . Additionally, GNNs also have the capability to learn embeddings for the entire graph, represented as h_G . In modern GNNs, a node's representation is updated iteratively by aggregating features from its neighbors, as originally described by Gilmer et al [43]. After k iterations, a node's representation reflects the structure of its k -hop neighborhood, allowing the GNN to capture both the local and global structural information present in the graph, making them a powerful tool for a wide range of

graph-based machine learning tasks. The k -th layer of a GNN can be expressed as follows:

$$h_v^{(k)} = \text{COMBINE}^{(k)}(h_v^{(k-1)}, \text{AGGREGATE}^{(k)}(\{(h_v^{(k-1)}, h_u^{(k-1)}, e_{uv}) : u \in N(v)\})) \quad (3.1)$$

where $h_v^{(k)}$ is the representation vector of node v at iteration/layer k , e_{uv} corresponds to the attributes of the edge connecting nodes u and v , and $N(v)$ designates the neighbor set of node v . The COMBINE function integrates information from the previous layer $k - 1$ and the aggregated features from neighboring nodes, adjusting the representation of node v accordingly. The AGGREGATE function sums up the information from the neighboring nodes, which is then utilized by the COMBINE function to update node representations. We initialize $h_v^{(0)} = X_v$. To obtain the overall representation of the entire graph, denoted as h_G , the READOUT function aggregates node features from the final iteration K as follows:

$$h_G = \text{READOUT}\left(\left\{h_v^{(K)} \mid v \in G\right\}\right) \quad (3.2)$$

The READOUT function in Graph Neural Networks, as highlighted by Ying et al. [190] and Zhang et al. [202], is designed to be permutation-invariant. This means it does not depend on the order of nodes and edges in the graph. Common techniques used for this purpose include averaging or more advanced graph-level pooling methods, which aggregate node features into a comprehensive representation of the entire graph. Indeed, numerous researchers have employed Graph Neural Network algorithms to model chemical structures [138, 175], given that molecules can be effectively represented as graphs [49].

In the context of our specific task, we represent a graph as $G = (V, E)$ with node attributes X_v for $v \in V$ and edge attributes e_{uv} for $(u, v) \in E$. Given a collection of graphs $\{G_1, G_2, \dots, G_n\}$ and their corresponding labels $\{y_1, y_2, \dots, y_n\}$, our objective is to learn a representation vector h_G for the entire graph that can predict the label $y_G = g(h_G)$. For instance, in the prediction of molecular properties, each G represents a molecular graph where nodes correspond to atoms, edges represent chemical bonds,

and the labels can signify various molecular properties or characteristics. While GNN models are proficient at capturing substantial information, there is a potential drawback of neglecting more fine-grained information of specific subunits within graphs, which could be pivotal for achieving precise predictions.

3.2.3 Self-attention

Self-attention is a fundamental mechanism in modern machine learning, widely applied in domains such as natural language processing and computer vision. Its strength lies in the ability to capture complex dependencies and relationships within an input sequence by allowing each element to attend to every other element. Through this process, the model learns to assign different levels of importance, or weights, to various parts of the sequence, thereby enabling a more nuanced understanding and more effective representation of contextual information.

The mechanism operates by projecting the input sequence into three distinct representations: query, key, and value vectors, which are obtained through learned linear transformations. The similarity between the query and key vectors is then computed, typically using a scaled dot-product, to determine attention scores that reflect the relative importance of different input components. These scores are subsequently used to compute a weighted sum of the value vectors, producing a context-aware representation of the input. This representation is further refined through additional layers, such as feed-forward neural networks and normalization steps, ultimately yielding the final output. By iteratively applying this process, self-attention empowers models to selectively focus on the most relevant information and to capture both local and long-range dependencies within the data, which is essential for tasks requiring complex pattern recognition and semantic understanding.

3.3 Methodologies

Our molecular description framework is composed of three complementary sub-modules: Fin2Vec, MG2Vec, and Ring2Vec. The Fin2Vec and MG2Vec components build upon established methodologies, drawing inspiration from fingerprint-based and molecular graph representations that have been widely reported in the literature. The key innovation

of our framework lies in the introduction of the Ring2Vec module, which specifically captures the structural and contextual information of molecular rings. When integrated with Fin2Vec and MG2Vec, the Ring2Vec module enables a more comprehensive molecular representation, thereby yielding substantial improvements in prediction accuracy and robustness compared to conventional approaches.

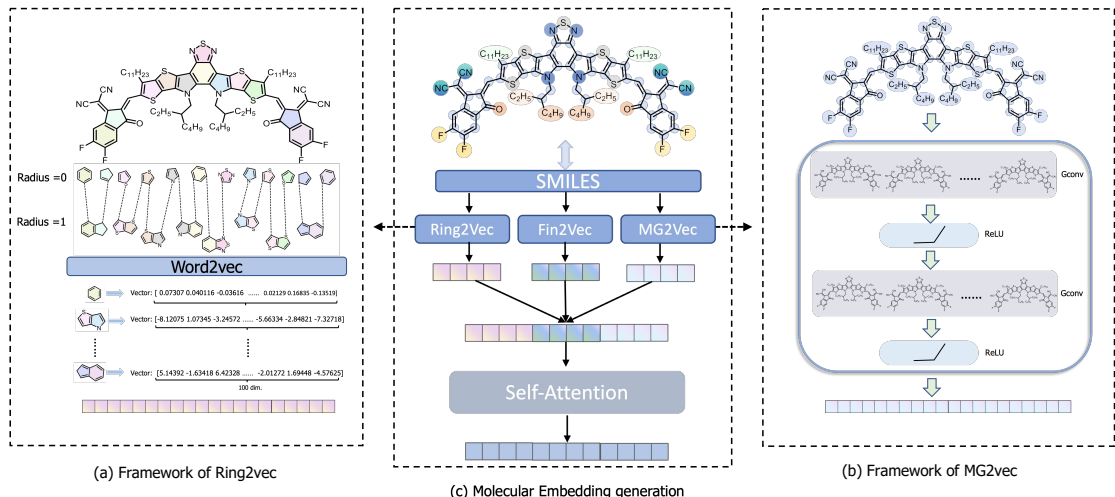


Figure 3.1: Overview of the proposed framework. (a) The Ring2Vec module begins by analyzing SMILES representations to detect ring-related substructures, defined as isolated rings or fused ring pairs. These structural tokens are then transformed into numerical representations using a Word2Vec-inspired embedding procedure, generating a vocabulary of several thousand ring embeddings. (b) The MG2Vec module processes molecular graphs with a GNN to capture atom-level features. (c) The Fin2Vec module generates embeddings from molecular fingerprints. Finally, embeddings from all three modules are integrated through a self-attention mechanism to construct a unified representation of NFA molecules.

Our method and computation work consists of the following three steps: 1. Pre-train a Ring2Vec model using the Harvard Clean Energy Project database, which enables the transformation of SMILES strings of NFA molecules into dense vector representations that effectively capture ring-level information. 2. Utilizing the commonly used self-attention mechanism [167], the three sets of vectors from Ring2Vec, Fin2Vec, and MG2Vec modules are fused into one set of high-dimensional representations that encompass atom-, subunit-level features. 3. These vectors are then used as the inputs to train various ML models (such as Random Forest, Extremely Randomized Trees,

etc.) to predict the properties of NFA molecules. For the prediction tasks, we employ a smaller, experimentally dataset derived from a previous publication that contains measured NFA properties, including HOMO and LUMO energy levels [115]. In contrast to our approach, the HOMO and LUMO levels reported in the Harvard CEP database are obtained through DFT calculations, whereas we prioritize experimental datasets to achieve higher reliability. (It should be noted that, in principle, HOMO/LUMO values are intrinsic molecular properties that can be directly measured using ultraviolet photoelectron spectroscopy (UPS). In practice, however, the smaller dataset used here, and indeed much of the OSC literature, derives HOMO/LUMO estimates from CV, where the redox potentials of NFA molecules in thin films are recorded. Consequently, the predictions associated with this dataset correspond to CV-based estimates of HOMO and LUMO, rather than UPS measurements.) Figure 3.1 provides a schematic illustration of the module integration framework together with the Ring2Vec component.

3.3.1 Ring2Vec Module

The Ring2Vec module is established via the following steps:

1. We first employ the RDKit cheminformatics toolkit to analyze the SMILES strings of organic solar cell molecules from the Harvard CEP database [83]. Unlike conventional fingerprint-based approaches, such as those adopted in Fin2Vec models, our method specifically targets the identification and extraction of all possible aromatic rings within the molecular structures. This ring-centric representation forms the foundation for the subsequent Ring2Vec training process, ensuring that chemically meaningful cyclic substructures are explicitly captured and utilized in our framework.
2. Once all aromatic rings are identified, we specify a collection of core features to describe them. These include the ring size, the number of single and double bonds, the presence of heteroatoms, fusion relationships with adjacent rings, atom types, and the electronegativity of heteroatoms. An algorithm is then designed to automatically assign numerical descriptors to each feature, enabling every ring to be encoded as a distinct numeric sequence. Accordingly, each OSC molecule

in the Harvard CEP dataset is characterized by several such sequences, with each sequence corresponding to one of its rings. This approach guarantees that every OSC molecule is represented through a unique set of ring-based numerical sequences.

3. Treating molecules as sentences and ring-derived subunits as words, we adapt algorithms from the classical Word2Vec framework to perform unsupervised pre-training on OSC molecules from the Harvard CEP database. The goal of this pretraining is to transform the simple numeric sequences associated with each molecule into high-dimensional embeddings that more effectively capture ring-specific characteristics. The outcome of this process resembles a vocabulary, in which each ring-based unit is assigned a unique high-dimensional vector representation. It is important to note that this pretraining stage does not incorporate the DFT-calculated HOMO/LUMO values available in the CEP database. Instead, the intent is to train a model capable of mapping SMILES strings directly into ring-based embeddings. Although the molecules in the Harvard CEP database differ structurally from NFAs, they share aromatic ring subunits as common building blocks. As a result, pretraining on the CEP database provides transferable knowledge that can be leveraged to improve property prediction for NFAs.

The pretraining procedure described in Step 3 can be elaborated as follows. When applying the Word2Vec model to molecular data, an essential step is to convert ring-based molecular structures into “sentences” suitable for training. In our methodology, this conversion can be achieved by treating either individual rings or pairs of adjacent rings as the fundamental structural units. This design is motivated by the observation that the electronic and structural properties of subunits are influenced not only by isolated rings but also by the interactions arising from neighboring rings. Consequently, for each molecule, there exist at least two potential sequence representations: one based on single-ring units and another based on double-ring (adjacent ring pair) units. In practice, these two strategies can be further diversified, yielding four distinct approaches for constructing ring-based molecular “sentences,” thereby enriching the diversity and robustness of the pretraining corpus:

Conversion 1: exclusively employs single-ring-based representation.

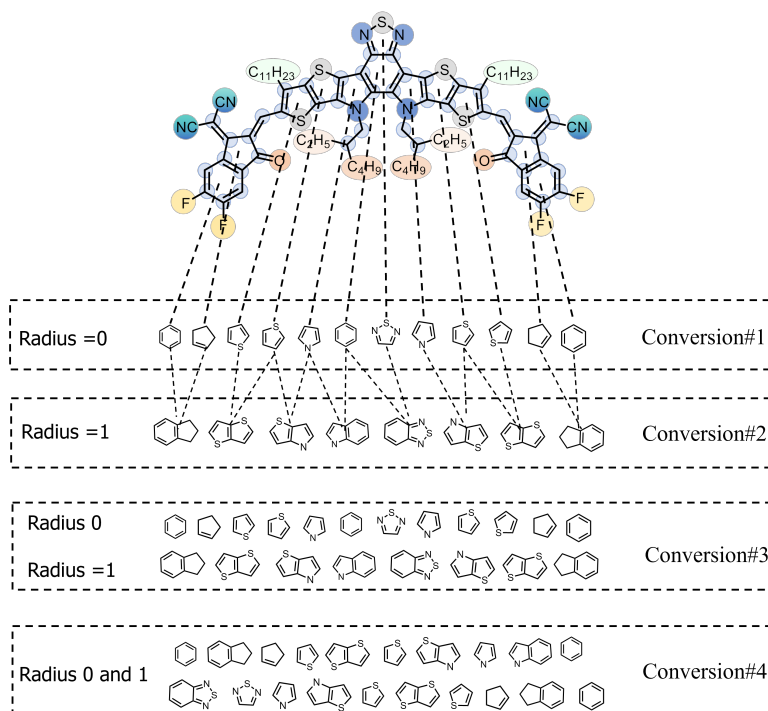


Figure 3.2: Depiction of the four distinct approaches used in Ring2Vec to convert molecular structures into sentence representations.

Conversion 2: exclusively uses representations based on two adjacent rings.

Conversion 3: describes the molecule by a series of single rings, followed by a series of combined adjacent rings.

Conversion 4: describes the molecule by alternating single rings and combined rings.

In addition to the different strategies for converting molecules into sequences, the Word2Vec framework itself provides two alternative algorithms for training embeddings: The first is the *Continuous Bag of Words* (CBOW) model, in which the task is to predict a target word based on its surrounding context words. In this approach, the order of context words is not considered, making it conceptually similar to the traditional bag-of-words model. The second is the *Skip-gram* model, which reverses the prediction task by using a target word to predict its surrounding context words. Skip-gram places greater emphasis on nearby words, thereby enabling the model to capture more fine-grained semantic relationships. In our pretraining process, we adopt and compare both CBOW and Skip-gram algorithms, in order to evaluate their relative effectiveness for learning ring-based molecular representations.

Following the pretraining process, we obtained the Ring2Vec model, which can be

viewed as a large vocabulary encompassing a diverse pool of single-ring and combined-ring units. For each of these ring-based units, the model is capable of generating high-dimensional vector embeddings that capture their structural and contextual properties. More importantly, when provided with the SMILES string of any OSC molecule, the Ring2Vec model can translate the molecular structure into a sequence of such embeddings. In this way, each OSC molecule is represented as a series of high-dimensional vectors, effectively encoding its ring-based features and preserving chemically meaningful molecular information.

3.3.2 *Fin2Vec Module*

In this module, we leverage the RDKit cheminformatics library to process SMILES (Simplified Molecular Input Line Entry System) strings, enabling the transformation of chemical structures into machine-readable formats suitable for advanced computational analysis. RDKit, a widely adopted open-source toolkit, provides a comprehensive suite of tools for cheminformatics, supporting tasks such as molecular visualization, property calculation, and structural manipulation. When applied to SMILES strings, compact, linear representations of molecular structures, RDKit offers robust functionality to parse these notations and convert them into detailed two-dimensional molecular graphs. These graphs explicitly capture critical chemical information, including atoms, bonds, and their connectivity, which are essential for subsequent computational workflows. Our workflow begins by extracting SMILES strings from structured CSV files, which serve as a common format for storing chemical data. Using RDKit's `Chem.MolFromSmiles` function, these strings are converted into molecular objects that encapsulate detailed atom and bond information, such as atomic numbers, bond types (e.g., single, double, or aromatic), and stereochemistry where applicable. These molecular objects form the foundation for further processing, enabling precise manipulation of chemical structures. Subsequently, we employ the `RDKFingerprint` function from RDKit to generate binary molecular fingerprints. These fingerprints are bit vectors where each bit corresponds to the presence or absence of specific substructural features or molecular patterns, such as functional groups or ring systems, derived from the molecular graph. The resulting fingerprints are fixed-size numerical representations that encode the struc-

tural and chemical properties of the molecules in a format suitable for computational tasks. These tasks include molecular similarity searches, clustering, and machine learning applications, such as quantitative structure-activity relationship (QSAR) modeling or virtual-Hungarian algorithm-based similarity searches. By transforming complex molecular structures into standardized, comparable vectors, the fingerprints facilitate quantitative comparisons across diverse molecules, enabling applications like virtual screening for drug discovery or chemical database analysis. Additionally, RDKit’s flexibility allows for further customization of fingerprints, such as adjusting bit vector length or incorporating specific feature sets, to meet the needs of specialized analyses. To ensure compatibility with downstream computational pipelines, the fingerprints are stored in a consistent format, typically as binary or integer arrays, which can be seamlessly integrated into machine learning frameworks or cheminformatics tools. This workflow not only streamlines the preprocessing of chemical data but also enhances the reproducibility and scalability of molecular analysis, making it a critical component of modern cheminformatics research and development.

3.3.3 MG2Vec Module

In this module, we utilize GNN embeddings to represent molecular structures in a high-dimensional feature space, offering a sophisticated approach to encoding the intricate details of molecular architectures. Unlike traditional molecular representations, GNN embeddings capture nuanced atom-level and bond-level information, enabling the creation of rich, context-aware feature vectors that reflect the unique chemical and structural characteristics of molecules. These embeddings serve as a powerful foundation for a wide array of molecular analysis tasks, including property prediction, molecular similarity assessment, and interaction modeling.

GNN embeddings are derived by processing molecular graphs, where atoms are represented as nodes and chemical bonds as edges. Each node is initialized with a feature vector that encapsulates properties such as atom type (e.g., carbon, oxygen, nitrogen), degree of connectivity, and other relevant chemical attributes. Similarly, edges are annotated with features like bond type (e.g., single, double, or aromatic) and bond order. The GNN model employs an iterative message-passing framework to update

these node feature vectors by aggregating information from neighboring nodes and their associated edges. This process captures both local and global structural patterns within the molecule, producing embeddings that reflect the molecular graph’s topology and chemical context.

The message-passing mechanism operates in several stages. Initially, each atom’s feature vector is updated by aggregating information from its immediate neighbors, considering factors such as neighboring atom types, bond characteristics, and structural connectivity. Aggregation functions, such as summation, mean, or attention-based mechanisms, combine this information into a cohesive representation. Subsequently, an update function transforms the aggregated data, refining the node’s feature vector to incorporate higher-order structural insights. This iterative process, typically repeated over multiple layers, allows the GNN to capture increasingly complex patterns, such as molecular substructures, functional groups, and long-range interactions between atoms.

To enhance the expressiveness of the embeddings, our approach integrates additional chemical properties into the GNN framework. For instance, we incorporate attributes like the electronegativity of heteroatoms into the initial node features. These chemically informed features enrich the embeddings, enabling them to capture not only structural but also physicochemical properties that are critical for downstream applications. After several iterations of message passing, the GNN produces a set of refined feature vectors for each atom, which collectively represent the molecule’s structural and chemical properties in a high-dimensional space.

To obtain a graph-level representation, we apply a READOUT function, as described by Ying et al. [190] and Zhang et al. [202], to aggregate the node-level embeddings into a single, fixed-size vector for the entire molecule. This READOUT function, often implemented as a sum, mean, or max-pooling operation, ensures that the final embedding encapsulates global molecular characteristics while preserving critical local details. The final output of the GNN model is a set of feature vectors, one for each atom in the molecule. These feature vectors serve as rich representations that encapsulate both local and global structural information, making them a powerful tool for molecular research. These GNN embeddings serve as a foundational component in our research, facilitating effective analysis and prediction of various molecular properties, including

molecular behavior, properties, and interactions. Their versatility and ability to capture detailed structural information make them a robust and indispensable component in the field of molecular research.

Table 3.1: Evaluation of predictive accuracy, measured by mean absolute error (MAE), across four distinct strategies for converting molecules into sentence-like representations.

	Methods	MAE ↓	MSE ↓	RMSE ↓	MAPE ↓	R^2 ↑
HOMO (eV)	Conversion#1	0.086 ±0.002	0.019 ±0.002	0.131 ±0.001	1.493 ±0.002	0.545 ±0.002
	Conversion#2	0.115 ±0.001	0.026 ±0.002	0.159 ±0.001	2.060 ±0.002	0.316 ±0.002
	Conversion#3	0.081 ±0.002	0.016 ±0.002	0.124 ±0.001	1.427 ±0.002	0.596 ±0.002
	Conversion#4	0.078 ±0.002	0.013 ±0.002	0.121 ±0.001	1.417 ±0.002	0.601 ±0.003
LUMO (eV)	Conversion#1	0.093 ±0.002	0.022 ±0.002	0.143 ±0.002	2.374 ±0.001	0.439 ±0.002
	Conversion#2	0.114 ±0.001	0.030 ±0.002	0.168 ±0.001	2.982 ±0.002	0.218 ±0.002
	Conversion#3	0.090 ±0.001	0.020 ±0.002	0.140 ±0.002	2.335 ±0.001	0.446 ±0.002
	Conversion#4	0.084 ±0.003	0.017 ±0.003	0.137 ±0.002	2.323 ±0.001	0.453 ±0.002
Eg (eV)	Conversion#1	0.105 ±0.002	0.031 ±0.002	0.169 ±0.002	5.897 ±0.003	0.618 ±0.002
	Conversion#2	0.133 ±0.001	0.042 ±0.002	0.199 ±0.001	7.668 ±0.002	0.469 ±0.002
	Conversion#3	0.099 ±0.004	0.027 ±0.001	0.161 ±0.002	5.660 ±0.002	0.654 ±0.002
	Conversion#4	0.094 ±0.004	0.024 ±0.002	0.156 ±0.002	5.650 ±0.003	0.660 ±0.002

3.4 Results

3.4.1 Prediction with only the Ring2Vec module.

In this section, we conduct prediction tasks targeting three key electronic properties of non-fullerene acceptors: the Highest Occupied Molecular Orbital (HOMO), the Lowest Unoccupied Molecular Orbital (LUMO), and the HOMO-LUMO gap (Eg). As described earlier, our workflow begins with pretraining the Ring2Vec model on the large-scale Harvard CEP database. Following this, we employ a second, smaller dataset to train machine learning models and carry out property predictions. This curated dataset comprises 1,318 data points collected from 558 published studies. The raw data are primarily derived from experimental measurements and include a range of important parameters, such as HOMO, LUMO, PCE, Eg, short-circuit current density (Jsc), open-circuit voltage (Voc), among others.

Table 3.2: Summary of mean absolute error (MAE) values from prediction tasks on HOMO, LUMO, and bandgap, evaluated using four molecule-to-sentence conversion methods combined with different ML models.

Properties	Methods	RF	ERT	GBRT	ANN	SVM
HOMO (eV)	Conversion#1	0.087	0.086	0.086	0.154	0.117
	Conversion#2	0.116	0.115	0.116	0.202	0.123
	Conversion#3	0.084	0.081	0.084	0.232	0.116
	Conversion#4	0.080	0.078	0.083	0.242	0.114
LUMO (eV)	Conversion#1	0.094	0.093	0.093	0.144	0.119
	Conversion#2	0.114	0.114	0.113	0.160	0.125
	Conversion#3	0.092	0.090	0.091	0.186	0.115
	Conversion#4	0.088	0.084	0.090	0.190	0.114
Eg (eV)	Conversion#1	0.105	0.105	0.106	0.137	0.145
	Conversion#2	0.136	0.133	0.133	0.166	0.144
	Conversion#3	0.104	0.099	0.102	0.170	0.139
	Conversion#4	0.099	0.094	0.101	0.161	0.136

Table 3.3: Summary of prediction results, reported as mean absolute error (MAE), for the Skip-gram and CBOW variants of the Word2Vec approach.

	Methods	MAE ↓	MSE ↓	RMSE ↓	MAPE ↓	R^2 ↑
HOMO (eV)	Skip-gram	0.078 ±0.002	0.013 ±0.002	0.121 ±0.001	1.417 ±0.002	0.601 ±0.003
	CBOW	0.081 ±0.002	0.016 ±0.001	0.124 ±0.001	1.423 ±0.001	0.593 ±0.003
LUMO (eV)	Skip-gram	0.084 ±0.003	0.017 ±0.003	0.137 ±0.002	2.323 ±0.001	0.453 ±0.002
	CBOW	0.089 ±0.002	0.022 ±0.002	0.142 ±0.002	2.339 ±0.001	0.443 ±0.003
Eg (eV)	Skip-gram	0.094 ±0.004	0.024 ±0.002	0.156 ±0.002	5.650 ±0.003	0.660 ±0.002
	CBOW	0.101 ±0.002	0.030 ±0.003	0.162 ±0.002	5.734 ±0.002	0.650 ±0.002

Table 3.4: MAE outcomes for predicting HOMO, LUMO, and bandgap are presented, based on Word2Vec embeddings derived from Skip-gram and CBOW variants. The corresponding values are listed in the table.

Properties	Methods	RF	ERT	GBRT	ANN	SVM
HOMO (eV)	Skip-gram	0.080	0.078	0.083	0.242	0.114
	CBOW	0.083	0.081	0.082	0.616	0.117
LUMO (eV)	Skip-gram	0.088	0.084	0.090	0.190	0.114
	CBOW	0.091	0.089	0.091	0.494	0.114
Eg (eV)	Skip-gram	0.099	0.094	0.101	0.161	0.136
	CBOW	0.105	0.101	0.102	0.369	0.131

In carrying out these regression tasks, we begin by evaluating the predictive performance of the Ring2Vec module in isolation, without incorporating the Fin2Vec or MG2Vec modules. This initial step allows us to systematically optimize the model parameters specific to Ring2Vec and assess its standalone effectiveness in capturing ring-level molecular features. Once the Ring2Vec module has been fully optimized, we integrate it with the Fin2Vec and MG2Vec modules to construct a more comprehensive representation. This combined framework leverages complementary information from atom-level, fingerprint-based, and ring-level descriptors, ultimately yielding improved prediction accuracy and more robust performance across the targeted molecular properties.

As outlined earlier, the Ring2Vec module supports four distinct conversion strategies for generating molecular representations, based on either single rings, double rings, or their combinations. To determine the most effective strategy, we first perform prediction tasks to evaluate the relative performance of these four conversion approaches. As shown in Table 3.1, we compare the performance of prediction tasks on HOMO/LUMO and Eg using the four molecule-to-sentence conversion methods under the Ring2Vec method, as described in earlier paragraphs and Figure 3.2. The ML model used in these prediction tasks is the ERT method, which is tested later to be the best among several ML models. We also performed similar prediction tasks using other ML models and summarized their performances in Table 3.2.

From the results in Table 3.1, it is evident that the fourth conversion strategy produces the lowest prediction errors for HOMO, LUMO, and Eg. This indicates that representing molecules with alternating single- and double-ring structures is the most effective option within the Ring2Vec module, and thus it is adopted as the default configuration in subsequent experiments. The outcome suggests that the electronic properties of NFAs are strongly influenced by the combined effects of both single and double rings. Moreover, the inferior performance of method 2 implies that relying solely on a double-ring representation is insufficient to provide an accurate molecular description.

Table 3.5: Prediction results benchmarked by MAE across molecular embeddings constructed from Fin2Vec, MG2Vec (GCN/GINE), Ring2Vec, and their combined feature space.

	Embedding	MAE ↓	MSE ↓	RMSE ↓	MAPE ↓	R^2 ↑
Fin2Vec	HOMO (eV)	0.103 ± 0.002	0.023 ± 0.001	0.152 ± 0.003	1.175 ± 0.002	0.414 ± 0.002
	LUMO (eV)	0.089 ± 0.002	0.021 ± 0.002	0.138 ± 0.001	2.293 ± 0.002	0.470 ± 0.001
	Eg (eV)	0.103 ± 0.001	0.030 ± 0.002	0.168 ± 0.001	5.843 ± 0.002	0.623 ± 0.002
GCN	HOMO (eV)	0.074 ± 0.002	0.016 ± 0.002	0.117 ± 0.002	1.268 ± 0.001	0.644 ± 0.003
	LUMO (eV)	0.074 ± 0.001	0.018 ± 0.003	0.121 ± 0.001	1.913 ± 0.001	0.594 ± 0.003
	Eg (eV)	0.086 ± 0.002	0.024 ± 0.004	0.142 ± 0.001	4.780 ± 0.003	0.735 ± 0.002
GINE	HOMO (eV)	0.083 ± 0.001	0.020 ± 0.002	0.136 ± 0.003	1.446 ± 0.001	0.527 ± 0.002
	LUMO (eV)	0.082 ± 0.002	0.017 ± 0.002	0.123 ± 0.001	2.114 ± 0.001	0.564 ± 0.001
	Eg (eV)	0.095 ± 0.002	0.026 ± 0.001	0.154 ± 0.000	5.229 ± 0.001	0.678 ± 0.001
Ring2Vec	HOMO (eV)	0.078 ± 0.002	0.013 ± 0.002	0.121 ± 0.001	1.417 ± 0.002	0.601 ± 0.003
	LUMO (eV)	0.084 ± 0.003	0.017 ± 0.003	0.137 ± 0.002	2.323 ± 0.001	0.453 ± 0.002
	Eg (eV)	0.094 ± 0.004	0.024 ± 0.002	0.156 ± 0.002	5.650 ± 0.003	0.660 ± 0.002
Fin2Vec + MG2Vec	HOMO (eV)	0.083 ± 0.002	0.020 ± 0.001	0.138 ± 0.002	1.437 ± 0.001	0.505 ± 0.001
	LUMO (eV)	0.080 ± 0.002	0.019 ± 0.001	0.133 ± 0.002	2.095 ± 0.002	0.520 ± 0.002
	Eg (eV)	0.095 ± 0.000	0.032 ± 0.001	0.174 ± 0.001	5.454 ± 0.000	0.579 ± 0.001
Ring2Vec + MG2Vec	HOMO (eV)	0.068 ± 0.001	0.011 ± 0.001	0.111 ± 0.001	1.222 ± 0.001	0.666 ± 0.001
	LUMO (eV)	0.070 ± 0.002	0.013 ± 0.002	0.118 ± 0.002	1.905 ± 0.002	0.593 ± 0.002
	Eg (eV)	0.075 ± 0.003	0.016 ± 0.002	0.130 ± 0.002	4.472 ± 0.003	0.766 ± 0.002
Ours (all three)	HOMO (eV)	0.063 ± 0.003	0.008 ± 0.003	0.103 ± 0.002	1.172 ± 0.002	0.710 ± 0.003
	LUMO (eV)	0.065 ± 0.002	0.009 ± 0.002	0.108 ± 0.003	1.176 ± 0.003	0.642 ± 0.003
	Eg (eV)	0.068 ± 0.004	0.011 ± 0.003	0.117 ± 0.004	4.175 ± 0.003	0.805 ± 0.004

In addition, we evaluated the Skip-gram and CBOW strategies, both originally introduced in the Word2Vec framework, to generate embeddings for NFA molecules.

Their predictive results are reported in Table 3.3, where the Extremely Randomized Trees model was employed as the predictor. For completeness, we also tested these embeddings with other machine learning models, and the corresponding outcomes are provided in Table 3.4. Since Skip-gram emphasizes sequential context while CBOW aggregates fragments in a “bag-of-words” manner, the superior results of Skip-gram indicate that the energetic properties of NFAs are strongly influenced by the ordering of rings. Therefore, Skip-gram is adopted as the default configuration in subsequent experiments. With these preliminary adjustments, we arrive at a well-optimized Ring2Vec module that can be effectively integrated with additional descriptors in later stages.

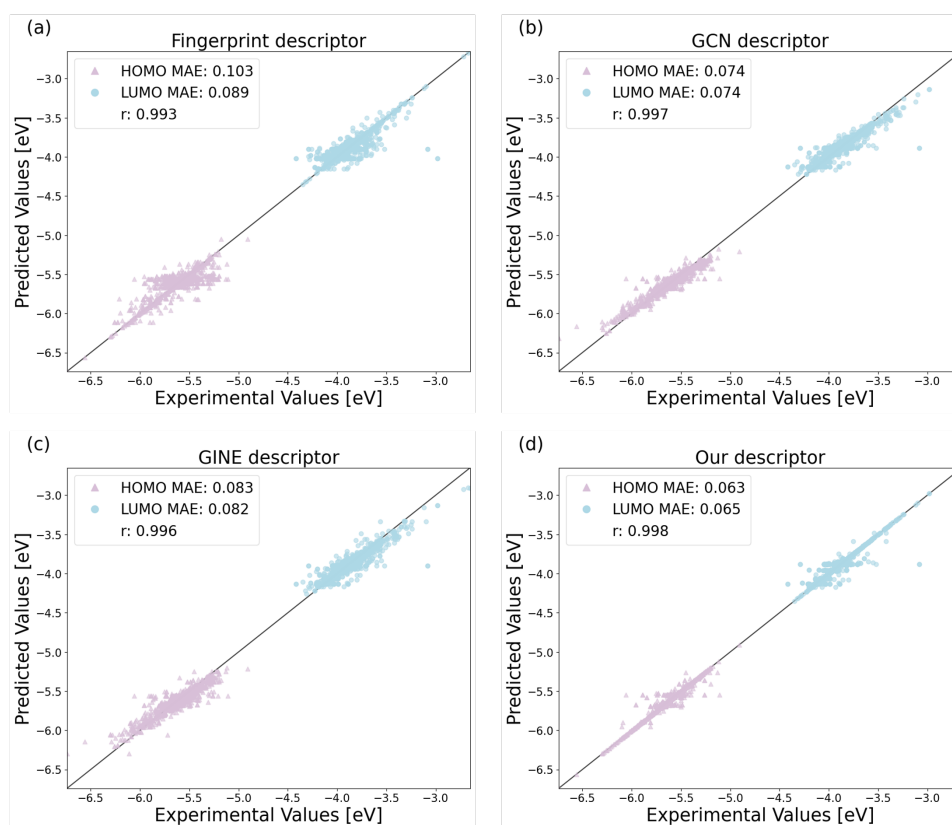


Figure 3.3: Experimental versus predicted HOMO and LUMO energy levels are shown using embeddings generated from (a) Fin2Vec, (b) MG2Vec, (c) Ring2Vec, and (d) their integrated representation (our proposed descriptor). In each plot, pink triangles indicate HOMO values and blue circles represent LUMO values. Predictions were obtained with the ERT model.

3.4.2 Comparison of prediction performance of different modules or combination of modules.

To benchmark the predictive capability of our Ring2Vec module against conventional Fin2Vec and MG2Vec approaches, we compared them with two widely used GNN baselines: the Graph Convolutional Network (GCN) and the Graph Isomorphism Network (GINE) [184, 75]. Table 3.5 reports the outcomes of this evaluation, covering seven distinct methods: (1) Fin2Vec, (2) GCN, (3) GINE, (4) Ring2Vec, (5) Fin2Vec + MG2Vec (GCN), (6) Ring2Vec + MG2Vec (GCN), and (7) the fully integrated framework, Ring2Vec + MG2Vec (GCN) + Fin2Vec, which achieves the strongest overall predictive accuracy.

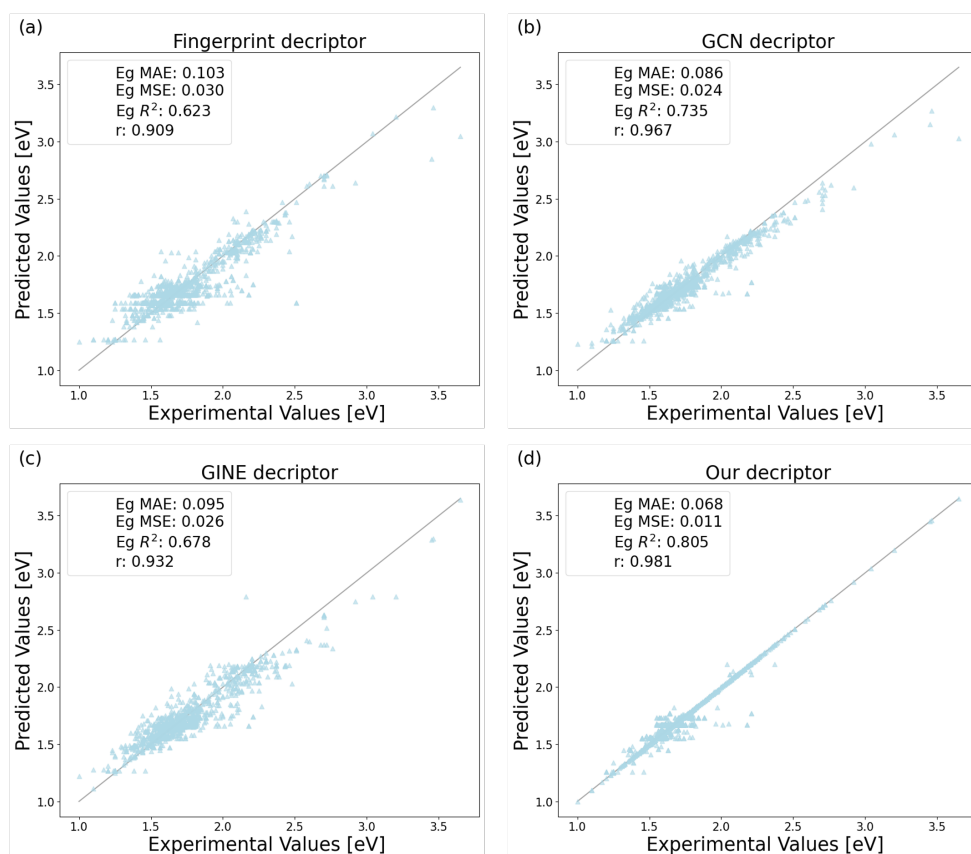


Figure 3.4: Experimental versus predicted bandgap (E_g) values derived from molecular embeddings of Fin2Vec, MG2Vec, Ring2Vec, and their integrated representation (our descriptor). Predictions were carried out with the ERT model.

Table 3.5, together with Figures 3.3 and 3.4, summarizes the predictive performance of different molecular representation strategies. The results clearly demonstrate the

advantage of our proposed method, which consistently yields the lowest error values across nearly all evaluation metrics. In particular, it achieves an MAE as low as 0.063 for HOMO prediction and attains the highest coefficient of determination ($R^2 = 0.805$) for bandgap Eg, highlighting its strong predictive capability. By contrast, the fingerprint-based approach produces noticeably larger errors, reflecting reduced accuracy. This weakness is especially pronounced for HOMO, where the fingerprint method reports an MAE of 0.103 and a much lower R^2 of 0.414, considerably worse than the other approaches.

Table 3.6: Summary of prediction performance, reported as MAE, MSE, RMSE, MAPE, R^2 , using the combined descriptor (Ring2Vec + Fin2Vec + MG2Vec) across different machine learning models.

		MAE ↓	MSE ↓	RMSE ↓	MAPE ↓	R^2 ↑
Random Forest (RF)	HOMO (eV)	0.069 ±0.003	0.010 ±0.003	0.109 ±0.003	3.287 ±0.004	0.671 ±0.004
	LUMO (eV)	0.071 ±0.003	0.013 ±0.001	0.113 ±0.004	4.601 ±0.003	0.615 ±0.004
	Eg (eV)	0.079 ±0.003	0.015 ±0.002	0.127 ±0.002	14.573 ±0.001	0.775 ±0.001
Extremely Randomized Trees (ERT)	HOMO (eV)	0.063 ±0.003	0.008 ±0.004	0.103 ±0.003	1.172 ±0.002	0.710 ±0.004
	LUMO (eV)	0.064 ±0.003	0.009 ±0.003	0.108 ±0.003	1.176 ±0.004	0.642 ±0.005
	Eg (eV)	0.068 ±0.004	0.011 ±0.003	0.117 ±0.004	4.175 ±0.003	0.805 ±0.004
Gradient Boosting Regression Trees (GBRT)	HOMO (eV)	0.067 ±0.002	0.010 ±0.002	0.105 ±0.003	3.387 ±0.000	0.692 ±0.001
	LUMO (eV)	0.069 ±0.002	0.011 ±0.002	0.112 ±0.003	4.704 ±0.003	0.623 ±0.001
	Eg (eV)	0.076 ±0.002	0.015 ±0.002	0.125 ±0.003	14.878 ±0.003	0.779 ±0.001
Artificial Neural Network (ANN)	HOMO (eV)	0.101 ±0.001	0.038 ±0.002	0.179 ±0.002	1.804 ±0.001	-0.004 ±0.001
	LUMO (eV)	0.096 ±0.001	0.021 ±0.001	0.145 ±0.002	2.555 ±0.001	0.381 ±0.001
	Eg (eV)	0.172 ±0.001	0.096 ±0.002	0.224 ±0.002	9.616 ±0.001	-0.100 ±0.001
Support Vector Machine (SVM)	HOMO (eV)	0.090 ±0.001	0.016 ±0.002	0.132 ±0.002	1.784 ±0.003	0.526 ±0.004
	LUMO (eV)	0.100 ±0.002	0.018 ±0.002	0.138 ±0.002	2.675 ±0.002	0.446 ±0.001
	Eg (eV)	0.120 ±0.001	0.023 ±0.001	0.152 ±0.003	7.055 ±0.002	0.679 ±0.001

The performance of GCN and GINE varies depending on the property under consideration. GCN surpasses the fingerprint baseline for HOMO and LUMO prediction, achieving MAE values of 0.074 for both. GINE, on the other hand, delivers more balanced results, with MAE scores of 0.083 for HOMO and 0.082 for LUMO. We also compare Ring2Vec and Ring2Vec+MG2Vec, where the latter provides further im-

provements in several metrics over the earlier models. Most notably, our proposed approach, integrating Ring2Vec, Fin2Vec, and MG2Vec, achieves the lowest prediction errors overall, as summarized in Table 3.5, confirming its superiority across molecular property prediction tasks.

In addition, Figure 3.3 presents the fitting results comparing experimental and predicted values of HOMO/LUMO. It is evident that in Figure 3.3d, which corresponds to our proposed descriptor, the data points are closely aligned along the $y = x$ diagonal, demonstrating excellent agreement between predictions and measurements. Moreover, the correlation coefficient (r) in Figure 3.3d reaches 0.998, the highest among the four subplots. These findings highlight the relative strengths and limitations of the different encoding strategies, while showing that our method achieves the highest level of accuracy and stability across multiple evaluation criteria. This emphasizes its strong potential for reliable molecular property prediction.

Next, we evaluate our molecular representation method that integrates the three modules (Ring2Vec, Fin2Vec, and MG2Vec), and summarize the results in Table 3.6. This combined descriptor is applied with several machine learning models, including random forest, extremely randomized trees, gradient boosted regression trees, artificial neural networks, and support vector machines, to predict key molecular properties: HOMO, LUMO, and Eg. Performance is assessed using standard regression metrics: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2). As reported in Table 3.6, the RF model delivers reasonable performance, reaching its best R^2 of 0.775 for Eg, but at the same time produces relatively larger errors (e.g., MAE = 0.079 for Eg) and records the highest MAPE (14.573%) among the tested models. By contrast, the ERT model achieves the strongest overall performance, with the highest R^2 (0.805) for Eg, reflecting superior fitting ability and predictive accuracy. This advantage is likely due to ERT’s randomized choice of split points and its use of the full sample during tree construction, which effectively lowers variance. In addition, ERT attains accurate predictions for HOMO and LUMO, with MAE values as low as 0.063 and 0.065, further demonstrating its strong estimation capability.

The GBRT model exhibits robust and competitive results, especially for Eg, where

it reaches an R^2 value of 0.779, ranking just behind the ERT model. By contrast, the ANN performs the worst, producing the largest prediction errors across all three properties, with Eg being most problematic (MAE = 0.172 eV, MSE = 0.096 eV). It even produces negative R^2 values for both HOMO and Eg, highlighting its poor ability to fit the data. The SVM shows a mixed pattern: it performs reasonably well for HOMO (MAE = 0.090 eV) but falls short on Eg, with a higher MAE of 0.120 eV and a lower R^2 of 0.679 compared with the other models.

3.5 Discussions

From the experimental outcomes, we observe that adopting an alternating single-double ring representation provides the most effective encoding strategy within the Ring2Vec module. This finding can be interpreted from several complementary perspectives. 1) *Structural complexity*: non-fullerene acceptors frequently display intricate molecular frameworks with diverse degrees of ring fusion, which exert strong influence on their electronic characteristics. Incorporating both single and double rings produces a more refined description that captures the interactions between adjacent fused units. 2) *Conjugation and electronic delocalization*: NFAs usually contain extended π -conjugated networks in which electrons are delocalized across successive single and double rings. By alternating the ring types in the representation, the conjugation effects can be modeled more faithfully, yielding a closer reflection of the actual electronic structure than when only one ring type is considered. This is especially important for predicting HOMO and LUMO levels, which are highly sensitive to conjugation length and continuity. 3) *Molecular orbital theory*: according to molecular orbital principles, overlapping p -orbitals in conjugated chains give rise to delocalized orbitals distributed throughout the molecule. An alternating representation can better approximate this orbital overlap, leading to stronger correlations with energy distributions and orbital alignment, which are essential for accurate estimation of the band gap.

In addition, we find that the Skip-gram approach consistently achieves stronger predictive performance than the CBOW method. While Skip-gram learns to predict the surrounding context from a single target token, CBOW estimates a target token based on its surrounding context. When applied to molecular structures, “tokens” may be interpreted as local structural units such as aromatic rings, whereas the “context” repre-

sents neighboring fragments connected to that unit. Because the positions and bonding patterns of adjacent fragments directly influence the overall molecular architecture, the Skip-gram formulation is better suited for capturing these dependencies. The superiority of Skip-gram can be attributed to several factors: 1) *Structural sensitivity*: Skip-gram emphasizes predicting neighboring fragments from a given unit, thereby capturing fine-grained structural relationships that are essential for modeling molecular “sentences.” Such sensitivity is especially useful for properties that depend on extended structural environments. 2) *Feature learning*: by learning contextual embeddings, Skip-gram effectively models how individual fragments contribute to molecular-level properties such as HOMO, LUMO, and Eg, allowing it to capture more detailed chemical semantics than CBOW. 3) *Scalability and flexibility*: Skip-gram typically scales more efficiently across datasets of varying sizes, delivering stable performance even under substantial changes in data volume. This robustness makes it a particularly attractive choice for molecular property prediction tasks, where dataset sizes often vary considerably.

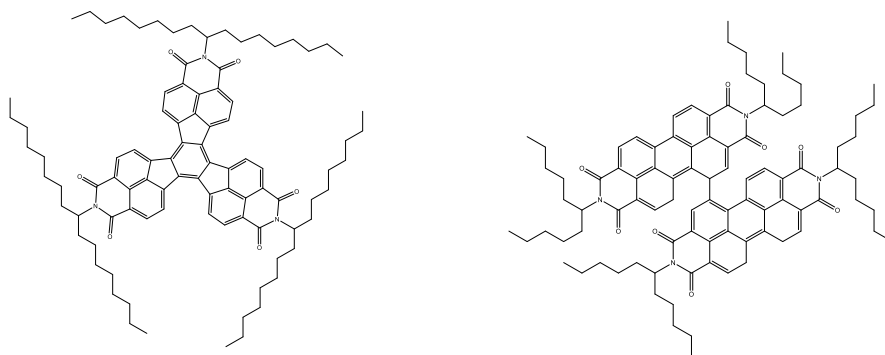


Figure 3.5: Outlier analysis showing the chemical structures of representative NFA molecules with comparatively large prediction deviations.

Furthermore, we conduct an outlier analysis to identify NFA molecules that exhibit relatively large prediction deviations and to examine the structural features responsible for such errors. As illustrated in Figure 3.5, two NFAs display prediction errors of approximately 0.15 eV for HOMO/LUMO levels, which is considerably higher than the average error of about 0.06 eV. Both of these molecules contain 2D-fused aromatic systems forming perylene-like core units. This type of two-dimensional fusion is distinct from the majority of NFAs (e.g., Y6 and ITIC), which are typically built from one-dimensional linearly fused aromatic chains. Because Ring2Vec is derived from Word2Vec, originally developed for sequential one-dimensional text data, it is naturally

more compatible with linear fused structures than with planar 2D-fused frameworks. From a practical standpoint, linear 1D aromatic architectures are not only more prevalent in existing NFA libraries but also easier to design and synthesize, which mitigates the significance of this limitation for most real-world applications. Nevertheless, this observation highlights a promising research direction: refining Ring2Vec to better handle complex 2D-fused aromatic systems, thereby expanding its applicability to a broader class of organic semiconductor materials.

The current Ring2Vec, Fin2Vec, and MG2Vec frameworks are limited in that they do not explicitly encode the three-dimensional (3D) geometry of NFA molecules. For most state-of-the-art NFAs, however, the conjugated cores remain nearly planar, meaning that the absence of explicit 3D information has only a minor influence on the prediction of HOMO and LUMO levels. The main molecular components exhibiting substantial 3D flexibility are the alkyl side chains. Yet, their contribution to electronic energy levels is relatively small compared to that of the conjugated backbone. Consequently, satisfactory predictive performance can still be achieved without a detailed account of 3D conformations. That said, for systems where the molecular backbone is significantly twisted or exhibits pronounced out-of-plane distortions, three-dimensional effects are expected to play a more critical role, and incorporating such information would become essential for accurate property prediction.

Overall, this study introduces a novel computational framework that markedly improves the prediction of NFA properties by adopting a multi-component strategy for molecular representation. The proposed approach integrates Ring2Vec, Fin2Vec, and MG2Vec modules to construct a more holistic description of molecular structures. Among these, the Ring2Vec module is particularly noteworthy for its Word2Vec-inspired design, which effectively captures subunit-level features such as aromatic rings-structural motifs that play a decisive role in shaping the electronic properties of NFAs used in OSCs. By moving beyond conventional atom-level descriptors and standard graph neural networks, this framework enables deeper understanding of the spatial organization and higher-order structural patterns governing NFA behavior. The advantages of the method are confirmed by experimental evaluation, where it surpasses both classical computational techniques and existing machine learning models. In particular, it achieves more accurate predictions of key OSC properties, including HOMO, LUMO, and Eg,

underscoring its ability to capture the inherent complexity of molecular systems. In addition, the automatic extraction of features directly from SMILES strings, without the need for manual annotation or domain-specific intervention, highlights its adaptability across diverse areas of materials research. Such capabilities have the potential to greatly accelerate the design and discovery of novel compounds, thereby supporting progress in organic solar cell technologies and beyond.

3.6 Experiments

Molecular Embedding Methods. We first provide additional implementation details for the molecular embedding approaches. Molecular fingerprints are represented as fixed-length binary vectors that encode structural and physicochemical features, enabling similarity comparisons between molecules based on bitwise operations. For graph-based encoders, we adopt the Graph Convolutional Network using PyTorch and PyTorch Geometric. The network is composed of stacked GCNConv layers, where node features are iteratively updated by aggregating information from their neighbors. Each convolutional block is followed by batch normalization and dropout to improve generalization and prevent overfitting. Global pooling is applied at the final stage to obtain graph-level embeddings, which are then used for downstream property prediction. The Graph Isomorphism Network Enhanced extends the standard GIN architecture by explicitly incorporating edge features into message passing. The GINEConv layer encodes both node and edge attributes, while multiple GINEConv layers are integrated through the GINet module to capture richer molecular representations.

Machine Learning Models. After pretraining the Ring2Vec model and the accompanying Fin2Vec and MG2Vec modules, we extract embeddings for non-fullerene acceptors from the experimental datasets. These embeddings serve as inputs to several machine learning algorithms for property prediction. The models evaluated include RF, ERF, GBRT, ANN, and SVM.

For tree-based models, key hyperparameters were tuned as follows: RF (#estimators = 80, max depth = 23, random state = 90), ERF (#estimators = 80, max depth = 23), GBRT (#estimators = 80, max depth = 20, random state = 90). The ANN consisted of a single hidden layer with 100 neurons, ReLU activation, and was trained with the

Adam optimizer (regularization coefficient $\alpha = 0.0001$). Default settings were retained for batch size, learning rate, and momentum, while random state was fixed at 90 for reproducibility. For the SVM, we employed an RBF kernel with $C = 1.0$ and $\epsilon = 0.2$, and standardized the input features before training.

Evaluation Protocol. To ensure reliability of the reported results, we adopted 10-fold cross-validation. In each run, the dataset was randomly partitioned into training and test subsets with a 9:1 ratio. Performance was quantified using commonly applied regression metrics: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2). The first four metrics evaluate predictive accuracy from different mathematical perspectives, while R^2 measures the explanatory power of the regression models with respect to the observed data.

$$\text{MAE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3.3)$$

$$\text{MSE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.4)$$

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3.5)$$

$$\text{MAPE}(y, \hat{y}) = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (3.6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.7)$$

where y_i and $\hat{y}_i$ are the actual and predicted values, $\bar{y}$ is the mean of the observed values, and m and n are the number of samples.

3.7 Conclusions

This study introduces a novel computational framework that advances the prediction of OSC properties by employing a multi-level representation strategy. The proposed

approach integrates Ring2Vec, Fin2Vec, and MG2Vec modules to deliver a richer characterization of molecular structures. Among these, the Ring2Vec module is particularly noteworthy for its Word2Vec-inspired design, which enables automated extraction of subunit-level semantics, especially for aromatic rings, a critical determinant of non-fullerene acceptor behavior. By moving beyond traditional atom-centric features and standard graph neural networks, this framework provides a more holistic perspective on the spatial and structural attributes that shape OSC performance. Empirical evaluation confirms its effectiveness: the model consistently surpasses conventional computational methods and prior machine learning baselines in both regression and classification tasks, achieving more precise predictions of essential OSC properties such as HOMO, LUMO, and PCE.

The framework also benefits from unsupervised pretraining on large-scale datasets, followed by fine-tuning with experimental measurements, forming a robust pipeline for predictive modeling in materials science. Leveraging resources such as CEPDB for pretraining, together with attention mechanisms for fusing diverse molecular descriptors, yields advanced molecular representations capable of capturing subtle structural complexity. In addition, the method’s automated extraction of features directly from SMILES strings, without manual curation or expert intervention, demonstrates its adaptability and broad applicability to material informatics. Such a capability has the potential to accelerate the design and discovery of new functional materials, thereby facilitating progress in organic solar technology.

In summary, this work marks a significant step forward in applying artificial intelligence and deep learning to computational chemistry, with a particular focus on OSC material development. The proposed framework not only achieves high predictive accuracy but also offers computational efficiency, establishing a new benchmark for property prediction pipelines and opening promising directions for renewable energy research.

Chapter 4: Ring-Enhanced Graph Transformer for PCE Prediction

This chapter introduces RingFormer, which is a three-tier graph-transformer for OSC molecules that unifies atom-level GINE message passing, ring-level cross-attention with a virtual molecule token, and bidirectional atom-ring communication, so that local chemistry, ring connectivity and their interplay are learned simultaneously without hand-crafted descriptors. Each layer first propagates bond information among atoms, then applies transformer-style global attention across aromatic rings-injecting ring-edge features directly into the attention keys and values-and finally fuses the two scales before stacking onward, yielding expressive, scale-aware embeddings. Evaluated on five benchmark datasets spanning 2.3 million DFT-generated structures (CEPDB) and several experimental sets (HOPV, PFD, NFA, PD), RingFormer outperforms 11 baselines including fingerprints, standard GNNs, pooling networks and other graph transformers: on CEPDB it lowers the MAE for PCE prediction to 0.189%, a 22.8% gain over the previous best, and it leads every competitor on multi-task benchmarks for HOMO, LUMO, band-gap, Voc and Jsc. As the first transformer to model explicit ring hierarchies in OSCs, RingFormer demonstrates that ring-aware attention sharply boosts accuracy while remaining dataset-agnostic.

This chapter is based on ” Z. Ding^{*}, **T. Zhang**^{*} (co-first author), Y. Li, J. Shi, and C. J. Zhang, ‘RingFormer: A Ring-Enhanced Graph Transformer for Organic Solar Cell Property Prediction’, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.”

4.1 Introduction

As the global demand for sustainable and renewable energy sources continues to grow, OSCs have garnered substantial attention within the scientific and industrial communities. Their unique capability to efficiently convert sunlight into electricity, combined with advantages such as mechanical flexibility, lightweight design, and potential for low-cost large-scale fabrication, distinguishes them from conventional photovoltaic technologies. These attributes not only position OSCs as a promising and eco-friendly alternative to fossil-fuel-based energy systems but also highlight their potential to play a critical role in addressing pressing global challenges, including climate change and the transition toward a sustainable energy economy [172]. Despite their significant potential as a renewable energy solution, the advancement of OSCs has been hindered by a predominant reliance on trial-and-error methodologies. These approaches often involve intricate molecular design, labor-intensive synthesis, and time-consuming device fabrication processes, which not only slow the pace of scientific discovery but also substantially increase research and production costs. Consequently, the efficiency and scalability of OSC development remain limited, underscoring the urgent need for more systematic, data-driven strategies to accelerate material discovery and device optimization [159]. To accelerate progress in OSC technology, there is growing enthusiasm for leveraging machine learning models to accurately predict the key properties of OSC molecules and device performance. By providing a transformative alternative to traditional experimental screening, ML-driven approaches offer the potential to streamline material discovery pipelines, reduce reliance on costly and time-consuming trial-and-error methods, and significantly expedite the design of efficient, high-performing OSCs. Moreover, the integration of ML with computational chemistry and experimental data can enable more systematic exploration of the vast molecular design space, paving the way for rational design strategies that were previously unattainable.

In this study, we aim to predict the properties of OSCs by analyzing the molecular structures of organic small-molecule semiconductors employed in their active layers. These molecules, functioning as either electron donors or acceptors, play a pivotal role in driving the photovoltaic effect. Their performance is intimately connected to their conjugated molecular frameworks, with structural features such as aromatic

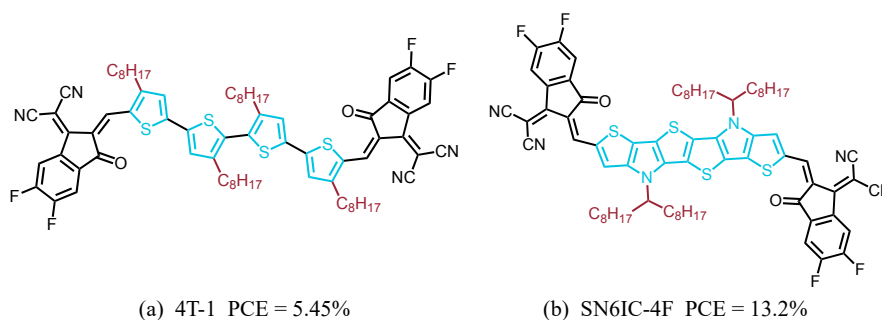


Figure 4.1: Illustrative examples of OSC molecules

rings and extended π -conjugation exerting a profound influence on their electronic and optical properties. These characteristics directly determine the efficiency of exciton generation, charge separation, and charge transport, thereby shaping the overall energy conversion efficiency of OSC devices [154]. In chemistry, a ring denotes a cyclic arrangement of atoms linked by covalent bonds, a structural motif that strongly affects molecular stability and electronic characteristics [72]. The architecture of ring systems, particularly those with intricate connectivity, has become a key theme in OSC material development [146], and may involve either fused or non-fused motifs [39]. As illustrated in Figure 4.1, the core of molecule 4T-1 (left, highlighted in yellow) consists of four thiophene rings that are non-fused and joined through single bonds, whereas SN6IC-4F (right) contains a fused S,N-heteroacene framework with six rings in which neighboring units share bonded atoms. Thus, both molecules display distinct ring topologies, forming extended conjugated frameworks of different complexity. Beyond the backbone, the presence of alkyl substituents (shown in blue) attached as side chains exerts a pronounced influence on performance. These groups regulate solubility, film morphology, and charge-transport efficiency, which are critical for achieving high photovoltaic output and overall device stability in OSC applications [116]. Such architectural contrasts between 4T-1 and SN6IC-4F ultimately produce markedly different power conversion efficiencies (PCEs). Consequently, reliable prediction of OSC properties requires models capable of simultaneously accounting for both the global ring-network architecture and the local chemical environment introduced by side chains.

Graph neural networks have emerged as a standard approach for molecular property prediction, as they naturally model molecules as graphs with atoms as nodes and chemical bonds as edges. This graph-based approach enables GNNs to effectively capture local atomic structures, such as functional groups and their interactions, which are

critical for understanding molecular behavior. However, despite their strengths, GNNs often struggle to adequately model higher-order structural patterns, such as the complex, interconnected ring systems found in OSC molecules, which are essential for accurately predicting properties like power conversion efficiency influenced by the global molecular architecture [53]. To overcome the shortcomings of purely atom-centered graph learning, researchers have proposed pooling strategies [38, 85] and motif-oriented frameworks [192, 196], which extend representations by incorporating higher-level molecular fragments. Despite these advances, such methods generally underrepresent the role of ring systems and the connectivity among them, features that are particularly critical for OSC molecules. This gap remains a major challenge, since conventional GNNs struggle to accurately capture cyclic architectures and long-range structural correlations within complex molecular graphs. As a result, their ability to encode higher-order patterns that strongly affect the electronic structure and photovoltaic response of OSC materials is limited [105, 19, 133], leading to reduced accuracy in property prediction tasks [34]. More recently, Zhu *et al.* [210] attempted to enrich GNNs with explicit ring descriptors. However, their framework still overlooks the essential inter-ring relationships. Consequently, existing graph-based models remain insufficient for faithfully characterizing the ring-dominated molecular frameworks of OSCs, thereby constraining predictive performance.

To bridge this gap, we propose RingFormer, a novel transformer-based graph architecture designed to characterize the complex structural organization of OSC molecules across both atomic and ring levels. As shown in Figure 4.2, RingFormer constructs a multi-level OSC graph that jointly represents individual atoms together with their aggregated ring frameworks. By adopting this hierarchical formulation, RingFormer is able to capture the interactions between local atomic features and extended conjugated ring domains, thereby establishing a solid basis for precise prediction of the electronic and photovoltaic behaviors that are critical for optimizing OSC device performance.

The hierarchical OSC graph is organized into three complementary layers: an atom-level representation, a ring-level representation, and an inter-level connector. At the atomic scale, the graph captures bonding patterns within the OSC molecule, highlighting local interactions and the connectivity of individual atoms. The ring-level representation then abstracts these details into larger cyclic units, modeling both fused

and non-fused rings to describe the extended conjugated frameworks that play a decisive role in OSC performance. Linking the two, the inter-level graph encodes the associations between atoms and rings, thereby preserving the multi-scale organization of the molecule. By combining information across these layers, the hierarchical OSC graph yields a refined and multi-dimensional description of molecular structure, supporting accurate prediction of critical properties, such as power conversion efficiency, that underpin progress in OSC technology.

RingFormer integrates localized message propagation with global attention to extract structural patterns across the three layers of the hierarchical OSC graph. At the atomic scale, it applies GNN-based message passing to capture fine-grained features such as bonds and functional groups, yielding detailed atom-level representations. At the ring scale, RingFormer introduces a cross-attention mechanism to model global dependencies within fused and non-fused ring assemblies that strongly influence OSC behavior. The inter-level component further enables interactions between atom and ring nodes, supporting an effective exchange of structural information. Each RingFormer layer concludes with a hierarchical fusion step, which unifies atom-, ring-, and inter-level information, allowing complementary insights from multiple structural resolutions. By stacking such layers, the model produces enriched graph embeddings that jointly encode atomic and ring-level characteristics of OSC molecules. To evaluate performance, we compared RingFormer with 11 established baselines across five diverse OSC datasets. The results show that RingFormer consistently delivers superior accuracy, capturing both local and global molecular features. On the large-scale CEPDB benchmark, it achieves a remarkable 22.77% relative gain over the strongest baseline, confirming its robustness and effectiveness in handling complex molecular architectures. In summary, our main contributions are:

- We address the key challenge of OSC property prediction and introduce RingFormer, the first transformer-based graph model that explicitly accounts for ring structures in OSC molecules.
- We design a hierarchical OSC graph representation that integrates atomic- and ring-level information, combining message passing with transformer-style attention to obtain rich molecular embeddings.
- We incorporate inter-level links between atoms and rings, together with a hi-

erarchical fusion module, to enhance the expressiveness and quality of learned representations.

- We conduct comprehensive experiments on five benchmark OSC datasets, demonstrating that RingFormer consistently achieves state-of-the-art predictive performance.

4.2 Related Work

OSCs have attracted substantial research interest as a highly promising technology for harnessing solar energy due to their potential for low-cost, flexible, and lightweight photovoltaic applications [34]. However, the process of screening candidate OSC molecules through laboratory experiments is both time-consuming and resource-intensive, posing significant challenges to rapid development [185]. To address these limitations, researchers have increasingly turned to machine learning methods to enable efficient and accurate prediction of OSC properties, thereby reducing the reliance on costly experimental approaches. Currently, fingerprint-based methods are widely used for OSC property prediction [34]. These methods generally rely on predefined molecular fingerprints, such as MACCS [32] and ECFP [137], as feature representations, which are subsequently used as inputs for conventional machine learning models including random forests and support vector machines. While these fingerprint-based methods are computationally efficient, they depend on simplified abstractions of molecular structures, which often overlook essential information and interactions, an issue particularly evident in OSC molecules with complex conjugated frameworks and ring systems [115]. This limitation hampers their ability to accurately predict properties essential for OSC performance, such as power conversion efficiency. Building on the demonstrated success of graph neural networks in drug discovery, Eibeck et al. [34] examined their potential for OSC property prediction. GNNs encode molecules as graphs, where nodes correspond to atoms and edges to chemical bonds, thereby capturing local atomic interactions. However, their findings revealed that conventional GNNs often underperform compared to fingerprint-based methods in OSC applications, achieving lower predictive accuracy. This is largely due to GNNs' limitations in capturing the higher-order structural patterns, such as complex ring systems and long-range dependencies, that are critical for OSC molecules. Consequently, the development of advanced, tailored models capable of

effectively predicting OSC properties remains a critical yet under-explored area in the pursuit of next-generation solar energy technologies.

Graph Representation Learning on Molecules. Given that molecules can be naturally represented as graphs, with atoms as nodes and chemical bonds as edges, graph neural networks [60, 210] have become a cornerstone for molecular property prediction across various applications. These models excel at capturing localized atomic interactions and structural features, making them well-suited for tasks like drug discovery. However, conventional GNNs face significant challenges in modeling high-order molecular structures, such as complex ring systems and long-range dependencies, which are critical for certain molecular properties [53]. In particular, GNNs struggle to adequately represent important molecular features like rings and their interconnections, which are pivotal in determining the behavior of OSC molecules [19, 105]. To overcome these limitations, researchers have explored advanced approaches, including graph pooling methods [38, 85], motif-based methods [192, 196], and graph transformers [138, 189, 74]. These approaches seek to model structural patterns that go beyond atom-level features, thereby offering a more holistic description of molecular architecture. Nevertheless, current methods remain limited in their treatment of complex ring assemblies and their interconnections, factors that are crucial for reliable prediction of OSC properties such as power conversion efficiency, which heavily depend on conjugated ring frameworks. To bridge this gap, we introduce RingFormer, a graph transformer architecture specifically developed to represent the ring, dominated structures of OSC molecules and to improve property prediction accuracy. RingFormer introduces a novel approach that explicitly captures both local atomic interactions and global ring-level patterns, ensuring a holistic representation of OSC molecular structures. By integrating a hierarchical graph structure and a tailored cross-attention mechanism, RingFormer effectively models the interplay between atoms, rings, and their interconnections, offering a robust solution to improve the accuracy and efficiency of OSC property prediction, thus advancing the development of next-generation solar energy technologies.

4.3 Problem Formulation

Data Model. In line with prior studies on OSC property modeling [34], we describe an OSC molecule using an atom-level graph $G_A = (V_A, E_A)$, where V_A denotes the collection of atom nodes and E_A represents the set of chemical bonds. Each node

$v_i \in V_A$ is associated with a feature vector $\mathbf{x}_{v_i} \in \mathbb{R}^{d_{V_A}}$, and each edge $e_{ij} \in E_A$ is assigned an attribute vector $\mathbf{e}_{ij} \in \mathbb{R}^{d_{E_A}}$, with d_{V_A} and d_{E_A} corresponding to the feature dimensions of nodes and edges, respectively.

Prediction Task. Given a molecular graph G_A , the objective is to learn a mapping $f : G_A \rightarrow y$, where y denotes the molecular property of interest. Typical targets include power conversion efficiency (PCE), highest occupied molecular orbital (HOMO), and lowest unoccupied molecular orbital (LUMO) [115]. Among these, PCE is of particular significance since it measures the ability of OSCs to convert solar energy into electricity [34]. Accordingly, this work emphasizes precise PCE prediction as its primary task.

4.4 Methodologies

Overview. To effectively represent both atomic- and ring-level characteristics of OSC molecules, we propose RingFormer, a framework that builds a hierarchical molecular graph and encodes it through specialized transformer layers for property prediction. As illustrated in Figure 4.2, this hierarchical graph integrates three components: atom-level, ring-level, and cross-level connections, providing a multi-scale view of OSC structures. The RingFormer layers jointly leverage message passing and attention mechanisms to refine node embeddings at each scale and to exchange information across levels. In the final stage, RingFormer combines atom- and ring-level embeddings into a unified representation, which is then used to predict target molecular properties of OSCs.

4.4.1 Hierarchical OSC Graph Construction

In this subsection, we describe the procedure for building the hierarchical OSC graph G that serves as the structural representation of an OSC molecule. As depicted in Figure 4.2, the ring system, defined by both ring types and their mutual connections, is a key determinant of the electronic and optoelectronic behavior of OSCs. A conventional flat molecular graph G_A , which encodes only atom-level structures, is often insufficient to capture such high-level semantics. To address this limitation, we introduce a hierarchical graph representation G that models an OSC molecule from three complementary levels.

Starting with the atom-level graph G_A , we construct a corresponding ring-level graph G_R that explicitly captures the composition and organization of ring systems.

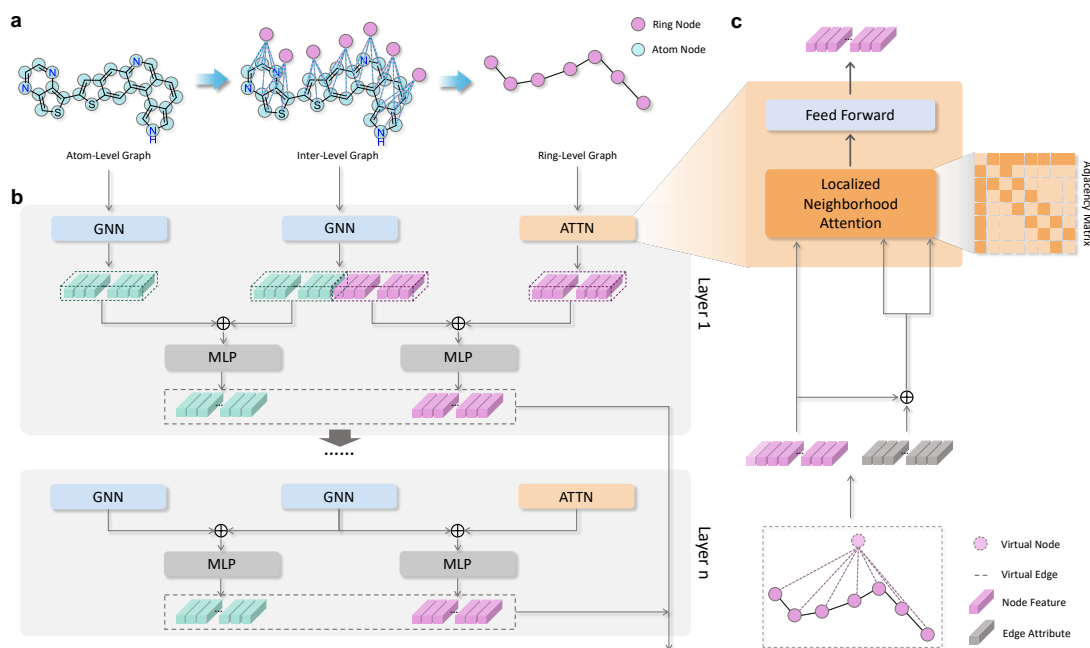


Figure 4.2: Overview of the RingFormer framework, shown here with two stacked RingFormer layers ($L=2$) for illustration.

Subsequently, we introduce a bipartite graph G_I , which connects atom nodes and ring nodes, thereby encoding the hierarchical relationships between local atomic structures and global ring motifs. By integrating these three components, we arrive at the complete hierarchical OSC graph $G = \{G_A, G_R, G_I\}$, which provides a comprehensive view of molecular architecture. In the following, we elaborate on the construction of each level in detail.

Ring-level Graph Construction. Given an input OSC molecule modeled as an atom-level molecular graph G_A (as introduced in the problem formulation section), the construction of the ring-level graph begins by systematically identifying and extracting the smallest rings from G_A . Each detected ring is then abstracted into a corresponding ring node, forming the set $V_R = \{v_1, v_2, \dots, v_{|V_R|}\}$. A *smallest ring* is formally defined as a cycle such that no proper subset of its constituent atoms can itself form a valid smaller ring, ensuring that only fundamental cyclic motifs are preserved without redundancy.

Rings in OSC molecules can vary widely in their atom composition, size, and aromaticity, all of which are critical to their electronic behavior. To capture these distinctions, each ring node $v_i \in V_R$ is represented by an attribute vector $\mathbf{x}_{v_i} \in \mathbb{R}^{d_{V_R}}$, where $\mathbf{x}_{v_i}$ encodes a one-hot representation of the ring type, and d_{V_R} denotes the total number of distinct ring categories under consideration.

Next, we introduce edges between pairs of ring nodes to reflect their structural

connectivity. Specifically, for two rings v_i and v_j , we add an edge e_{ij} to the set E_R if either of the following conditions is satisfied: (1) the two rings share one or more atoms, thereby indicating fused or overlapping ring systems, or (2) the two rings are bridged by a linear chain composed of one or more non-aromatic bonds, which often appear as linkers between conjugated motifs. Each edge $e_{ij} \in E_R$ is further associated with an attribute vector $\mathbf{e}_{ij} \in \mathbb{R}^{d_{E_R}}$, represented as a one-hot encoding of the connection type. Here, d_{E_R} specifies the total number of defined connection categories. The classification of connection types is determined by the precise structural relationship: for condition (1), by the identity and count of atoms shared between rings, and for condition (2), by the atomic and bonding composition of the intervening chain.

By assembling the set of ring nodes V_R together with their interconnections E_R , we obtain the ring-level molecular graph $G_R = (V_R, E_R)$. This representation provides a higher-level abstraction of the OSC molecule, emphasizing the interplay of aromatic substructures and their linkages, which play a decisive role in charge transport, conjugation, and overall optoelectronic properties.

Inter-level Graph Construction. To explicitly capture the hierarchical relationships between atoms and rings, we further construct an inter-level graph G_I that links the atom-level graph G_A and the ring-level graph G_R . Formally, for a ring node $v_i \in V_R$ and an atom node $v_j \in V_A$, we introduce an inter-level edge $e_{ij} \in E_I$ if the atom node v_j belongs to the set of constituent atoms represented by the ring node v_i . In this manner, G_I encodes the membership relation between atoms and rings, yielding an inter-level graph defined as $G_I = (V_A \cup V_R, E_I)$. Since G_I is treated as an undirected graph, it enables bidirectional message propagation between any pair of connected atom and ring nodes, thereby allowing atom-level features to inform ring-level representations and vice versa.

Finally, we integrate the three levels of graphs into a unified hierarchical OSC graph, denoted as $G = \{G_A, G_R, G_I\}$. This hierarchical representation comprehensively encapsulates the multiscale structural characteristics of OSC molecules, from local atomic interactions to high-level ring connectivity and cross-level membership relations. By doing so, it provides a principled foundation for accurate molecular property prediction, enabling models to exploit structural information across different resolutions.

4.4.2 RingFormer Layer

In this section, we describe the RingFormer layers designed to learn highly expressive representations for both atom and ring nodes within the hierarchical OSC graph G . Given the hierarchical structure of G , which encompasses atom-level, ring-level, and inter-level graphs, RingFormer layers integrate the strengths of message-passing and global attention mechanisms to effectively capture the distinct structural patterns at each level. As depicted in Figure 4.2, each RingFormer layer comprises three key components: an atom-level message-passing module, a ring-level cross-attention module, and an inter-level message-passing module. The atom-level module employs message-passing graph neural networks to encode localized atomic interactions, while the ring-level module utilizes a novel cross-attention mechanism to model global patterns and interconnections within the ring system. The inter-level module facilitates information exchange between atom and ring nodes, capturing their hierarchical relationships. At the conclusion of each RingFormer layer, a hierarchical fusion strategy integrates the latent information from all three levels to produce comprehensive and expressive node representations for both atoms and rings. These representations are then propagated to the subsequent layer for further refinement. In the following subsections, we provide a detailed explanation of each module and its role in enhancing the predictive capability of the RingFormer framework.

Atom-level Message Passing Module. On the G_A , we aim to learn atom node representations that capture the local chemical structures in OSC molecules, such as function groups. As GNNs are good at mining patterns in local structures [53], we adopt GNNs to perform message passing on G_A . In the l -th RingFormer layer, GNNs are applied on the G_A as follows:

$$\mathbf{h}_{v_i,A}^{(l)} = \text{GNN}_A^{(l)}(\mathbf{h}_{v_i}^{(l-1)}, \{\mathbf{h}_{v_j}^{(l-1)} | v_j \in N_{G_A}(v_i)\}, \{\mathbf{e}_{ij}\}), \quad (4.1)$$

where $v_i, v_j \in V_A$ denote nodes in G_A , $N_{G_A}(v_i)$ is the neighborhood of node v_i in G_A , $\mathbf{e}_{ij}$ is the edge attribute of edge $e_{ij} \in E_A$, and $\mathbf{h}_{v_i}^{(l-1)} \in \mathbb{R}^d$ is the node representation of node v_i from the preceding RingFormer layer. Specially, we have $\mathbf{h}_{v_i}^{(0)} = \mathbf{W}_a \mathbf{x}_{v_i}$, where $\mathbf{W}_a \in \mathbb{R}^{d \times d_{V_A}}$ are learnable parameters to transform node attribute $\mathbf{x}_{v_i}$, d is the hidden dimension. Here, $\text{GNN}_A^{(l)}$ can be arbitrary GNN architecture. Throughout the paper, we use GINE [60] as the default GNN backbone. We will also study the effect of different GNN backbones later. $\mathbf{h}_{v_i,A}^{(l)} \in \mathbb{R}^d$ is the output node representation of the atom-level message passing module in the l -th RingFormer

layer.

Ring-level Cross-attention Module. In encoding the ring-level graph G_R , our objective is to learn ring node representations that effectively capture the global structure of the entire ring system. To this end, we propose a novel transformer-based *cross-attention module* specifically designed for G_R . Graph transformers [189, 80] have recently demonstrated strong advantages over conventional GNNs in modeling long-range node dependencies and uncovering structural patterns that extend beyond localized neighborhoods [133]. These properties make them particularly suitable for ring-level modeling, where capturing interactions across distant cyclic motifs is crucial for representing the conjugated architectures of OSC molecules.

However, existing graph transformers still face notable limitations when applied to the chemically semantic-rich structures of ring-level graphs in OSC molecules. In particular, a primary shortcoming lies in their inadequate utilization of edge attributes, which encode essential information about how two rings are interconnected in G_R . Conventional graph transformers typically compute self-attention with node representations serving as queries, keys, and values, while edge attributes are only incorporated as auxiliary attention biases. As a result, the valuable chemical semantics carried by edge attributes in G_R are not fully integrated into the final ring node representations, limiting the model’s ability to capture meaningful structural relationships among rings.

Moreover, graph transformers also suffer from efficiency issues when incorporating edge attributes. For example, Graphormer employs fully-connected attention over all node pairs in a graph and relies on shortest-path distances as edge attributes. This design requires computing path information between every possible node pair, which introduces substantial computational overhead and scales poorly with the number of nodes [151].

To overcome the aforementioned limitations, we design a *ring-level cross-attention module* that incorporates two key components: (1) a cross-attention mechanism and (2) a localized neighborhood attention strategy. These designs enable more effective utilization of edge attributes while preserving computational efficiency.

First, to better exploit edge information, we replace the conventional node-centric *self-attention* mechanism in graph transformers with a *cross-attention* strategy. Formally, for a source ring node $v_j \in V_R$ connected to a target ring node $v_i \in V_R$ through an edge $e_{ij} \in E_R$, we treat the representation of v_i as the query, while the combination of the representation of v_j and the associated edge attribute $\mathbf{e}_{ij}$ serves as the key and value. In this way, edge attributes contribute not only to the computation of attention scores but also directly to the aggregated target node representation.

Subsequently, in the computation of multi-head attention, the queries, keys, and values for the c -th attention head in the l -th layer are defined as:

$$\begin{aligned} \mathbf{z}_{ij}^{(l)} &= \text{MLP}_Z^{(l)}(\mathbf{h}_{v_j}^{(l-1)} || \mathbf{e}_{ij}), \\ \mathbf{q}_{i,c}^{(l)} &= \mathbf{W}_{q,c}^{(l)} \mathbf{h}_{v_i}^{(l-1)}, \mathbf{k}_{ij,c}^{(l)} = \mathbf{W}_{k,c}^{(l)} \mathbf{z}_{ij}^{(l)}, \mathbf{v}_{ij,c}^{(l)} = \mathbf{W}_{v,c}^{(l)} \mathbf{z}_{ij}^{(l)}, \end{aligned} \quad (4.2)$$

where $||$ denotes the concatenation operation, and $\text{MLP}_Z^{(l)}$ represents a multi-layer perceptron (MLP) with a single hidden layer. The matrices $\mathbf{W}_{q,c}^{(l)}, \mathbf{W}_{k,c}^{(l)}, \mathbf{W}_{v,c}^{(l)} \in \mathbb{R}^{\frac{d}{C} \times d}$ are learnable projection parameters for the c -th attention head, and C denotes the total number of attention heads. In particular, the initial node embedding is defined as $\mathbf{h}_{v_i}^{(0)} = \mathbf{W}_r \mathbf{x}_{v_i}$, where $\mathbf{W}_r \in \mathbb{R}^{d \times d_{v_R}}$ is a learnable transformation matrix that maps the raw node attribute $\mathbf{x}_{v_i}$ into the d -dimensional embedding space.

Furthermore, to alleviate the computational burden, we draw inspiration from [151] and replace the fully-connected graph attention with a localized neighborhood attention mechanism augmented by a virtual molecule node. Concretely, in the ring-level cross-attention module, we introduce a virtual molecule node into G_R and connect it to all ring nodes through virtual edges. This design allows global information to be aggregated and propagated via the virtual node, thereby avoiding the need to compute attention across all node pairs. The representation of the virtual node is initialized with a learnable embedding, while the attributes of the virtual edges are encoded as one-hot vectors distinct from those of real edges in G_R . Subsequently, localized neighborhood attention is computed between each target node and its immediate neighbors as follows:

$$\alpha_{ij} = \frac{\langle \mathbf{q}_{i,c}^{(l)}, \mathbf{k}_{ij,c}^{(l)} \rangle}{\sum_{v_k \in N_{G_R}(v_i)} \langle \mathbf{q}_{i,c}^{(l)}, \mathbf{k}_{ik,c}^{(l)} \rangle}, \quad (4.3)$$

where $\mathbf{q}_{i,c}^{(l)}, \mathbf{k}_{ij,c}^{(l)}, \mathbf{v}_{ij,c}^{(l)} \in \mathbb{R}^{\frac{d}{C}}$ are the query, key, value vectors calculated in Eq.(4.2), $N_{G_R}(v_i)$ is the neighborhood of node v_i in G_R , and $\langle \mathbf{q}, \mathbf{k} \rangle = \frac{\mathbf{q}^T \mathbf{k}}{\sqrt{d}}$. Then, we update ring node representations based on cross-attention as follows:

$$\hat{\mathbf{h}}_{v_i,R}^{(l)} = \mathbf{W}_s^{(l)} \mathbf{h}_{v_i}^{(l-1)} + \mathbf{W}_o^{(l)} \left\| \sum_{c=1}^C \left(\sum_{v_j \in N_{G_R}(v_i)} \alpha_{ij} \mathbf{v}_{ij,c}^{(l)} \right) \right\|, \quad (4.4)$$

where $\mathbf{W}_s^{(l)}, \mathbf{W}_o^{(l)} \in \mathbb{R}^{d \times d}$ are learnable projection matrices, and $\hat{\mathbf{h}}_{v_i,R}^{(l)} \in \mathbb{R}^d$ denotes the output representation of node v_i after the multi-head cross-attention operation. In particular, the initial node embedding is computed as:

$$\mathbf{h}_{v_i}^{(0)} = \mathbf{W}_r \mathbf{x}_{v_i} || \mathbf{p}_{v_i}, \quad (4.5)$$

where $\mathbf{p}_{v_i} \in \mathbb{R}^{d_p}$ is a real-valued embedding vector of dimension d_p that serves as a degree-based positional encoding [189]. Here, $\mathbf{W}_r \in \mathbb{R}^{(d-d_p) \times d_{v_R}}$ is a learnable transformation matrix that maps the ring-type attribute $\mathbf{x}_{v_i}$ into a $(d - d_p)$ -dimensional subspace, which is then concatenated with the positional encoding to form the complete d -dimensional initialization for v_i .

Following the standard design of transformer blocks [167], the node representations obtained from the cross-attention module are subsequently passed through a feed-forward network, and the updated representations are computed as follows:

$$\mathbf{h}_{v_i,R}^{(l)} = \text{FFN}^{(l)}(\hat{\mathbf{h}}_{v_i,R}^{(l)} + \mathbf{h}_{v_i}^{(l-1)}) \quad (4.6)$$

where $\text{FFN}^{(l)}$ is the feed-forward block in the l -th layer, and $\mathbf{h}_{v_i,R}^{(l)} \in \mathbb{R}^d$ is the output node representation of ring-level cross-attention module in the l -th RingFormer layer.

Inter-level Message Passing Module. In the inter-level message passing module, information is exchanged between atom and ring nodes on the bipartite graph G_I , enabling atom representations to capture the global context of the higher-level ring system, while ring representations are simultaneously enriched by the local structural details of their constituent atoms. Formally, GNNs are employed within this module to perform the inter-level message passing, which is defined as follows:

$$\mathbf{h}_{v_i,I}^{(l)} = \text{GNN}_I^{(l)}(\mathbf{h}_{v_i}^{(l-1)}, \{\mathbf{h}_{v_j}^{(l-1)} | v_j \in N_{G_I}(v_i)\}), \quad (4.7)$$

where $v_i, v_j \in V_I$ denotes nodes in G_I , $N_{G_I}(v_i)$ is the neighborhood of node v_i in G_I , and $\mathbf{h}_{v_i,I}^{(l)} \in \mathbb{R}^d$ is the output node representation of inter-level message passing module in the l -th RingFormer layer.

Hierarchical Messages Fusion. After learning node representations across the three levels of graphs, the l -th RingFormer layer produces two types of embeddings for each atom and ring node, corresponding to different graph hierarchies. Specifically, for an atom node $v_i \in V_A$, the representation $\mathbf{h}_{v_i,A}^{(l)}$ is obtained from the atom-level message passing module on G_A , while $\mathbf{h}_{v_i,I}^{(l)}$ is derived from the inter-level message passing module on G_I . Similarly, for a ring node $v_j \in V_R$, the embedding $\mathbf{h}_{v_j,R}^{(l)}$ is produced by the ring-level cross-attention module on G_R , and $\mathbf{h}_{v_j,I}^{(l)}$ is obtained through the inter-level message passing module on G_I .

To enable effective information fusion across these hierarchical representations, the two embeddings associated with each node are combined as follows:

$$\mathbf{h}_{v_i}^{(l)} = \begin{cases} \text{MLP}_A^{(l)}(\mathbf{h}_{v_i,A}^{(l)} || \mathbf{h}_{v_i,I}^{(l)}) & \text{if } v_i \in V_A \\ \text{MLP}_R^{(l)}(\mathbf{h}_{v_i,R}^{(l)} || \mathbf{h}_{v_i,I}^{(l)}) & \text{if } v_i \in V_R \end{cases} \quad (4.8)$$

where $\mathbf{h}_{v_i}^{(l)} \in \mathbb{R}^d$ is the final node representation output by the l -th RingFormer layer for node $v_i \in V_A \cup V_R$. $\mathbf{h}_{v_i}^{(l)}$ is then sent to the next RingFormer layer for further update.

4.4.3 Prediction Layer

After stacking L RingFormer layers, we aggregate the output representations of all nodes to construct a comprehensive graph-level embedding for the hierarchical OSC graph. Concretely, for each node $v_i \in V_A \cup V_R$, where V_A denotes the set of atom nodes and V_R denotes the set of ring nodes, we concatenate its representations obtained from all L RingFormer layers to derive the final node embedding. This concatenation strategy integrates information learned at multiple depths, thereby capturing structural patterns across different scales within the hierarchy. As a result, the final node representations jointly encode both local atomic interactions and global ring-system characteristics, providing a robust and holistic foundation for accurate prediction of OSC molecular properties.

$$\mathbf{h}_{v_i} = \text{CONCAT}(\mathbf{h}_{v_i}^{(0)}, \mathbf{h}_{v_i}^{(1)}, \dots, \mathbf{h}_{v_i}^{(L)}), \quad (4.9)$$

where CONCAT denotes the concatenation operation, and $\mathbf{h}_{v_i} \in \mathbb{R}^{d \times (L+1)}$ represents the concatenated embedding of node v_i . Subsequently, we separately aggregate the node representations of atom nodes and ring nodes in G using sum pooling, and then concatenate the two pooled results to obtain the final graph-level representation:

$$\mathbf{h}_G = \text{POOL}(\{\mathbf{h}_{v_i} | v_i \in V_A\}) || \text{POOL}(\{\mathbf{h}_{v_i} | v_i \in V_R\}), \quad (4.10)$$

where POOL denotes the sum pooling operation, and $\mathbf{h}_G \in \mathbb{R}^{2d \times (L+1)}$ is the final graph representation of the hierarchical OSC graph G . Since the target properties of OSC molecules in this work are continuous values, the prediction task is formulated as a regression problem. Accordingly, we project the graph representation into the output logits $\hat{y}_G$ using a single linear layer. Given a batch of training molecules, RingFormer is optimized using the mean absolute error (MAE) loss, defined as:

$$L = \frac{1}{B} \sum_{b=1}^B |y_{G_b} - \hat{y}_{G_b}|, \quad (4.11)$$

where B is the batch size, and $\hat{y}_{Gb}$ and y_{Gb} represent the predicted and ground-truth molecule properties, respectively.

4.5 Results

4.5.1 Experimental Settings

Datasets and Evaluation Metrics. To assess the performance of our property prediction framework, we curated five distinct OSC molecule datasets, as summarized in Table 4.1. Notably, the CEPDB dataset [50] is derived using density functional theory (DFT) calculations, providing a computationally generated benchmark for OSC molecular properties. The remaining four datasets, HOPV [104], PFD [119], NFA [115], and PD [115], comprise experimentally validated OSC molecules, each representing diverse molecular types with properties critical for photovoltaic applications. To ensure robust evaluation, we adopted the scaffold-based splitting method [181], a standard approach in molecular machine learning, to partition each dataset into training, validation, and testing sets with a ratio of 6:2:2. This splitting strategy groups molecules by their structural scaffolds, promoting generalizability by ensuring that similar molecular backbones are not split across sets. Following established practices in the field [34, 141], we employ MAE as the primary evaluation metric to quantify the accuracy of property predictions across all datasets, enabling a consistent comparison with prior studies.

Dataset Details

CEPDB: The Clean Energy Project Database [50], curated from Harvard University, comprises approximately 2.3 million organic semiconductor molecules identified as potential donor materials for OSCs. Each molecule in the dataset is represented by a SMILES string, and their OSC-related properties are computed using DFT methods, ensuring high-fidelity theoretical predictions. During the curation process, we utilized RDKit¹ to filter out invalid SMILES strings, resulting in a refined dataset of 2,225,974 valid molecules. The dataset includes six key properties critical for OSC property prediction: power conversion efficiency, highest occupied molecular orbital, lowest unoccupied molecular orbital, band gap, open circuit voltage, and short-circuit current density. To facilitate graph-based modeling, the SMILES strings are converted into atom-level molecular graphs using PyTorch-Geometric², enabling seamless integration with graph neural network frameworks for accurate and efficient property prediction.

HOPV: The Harvard Organic Photovoltaic (HOPV) dataset [104] contains 350 carefully

¹<https://www.rdkit.org/docs/index.html>

²<https://pytorch-geometric.readthedocs.io/en/latest/index.html>

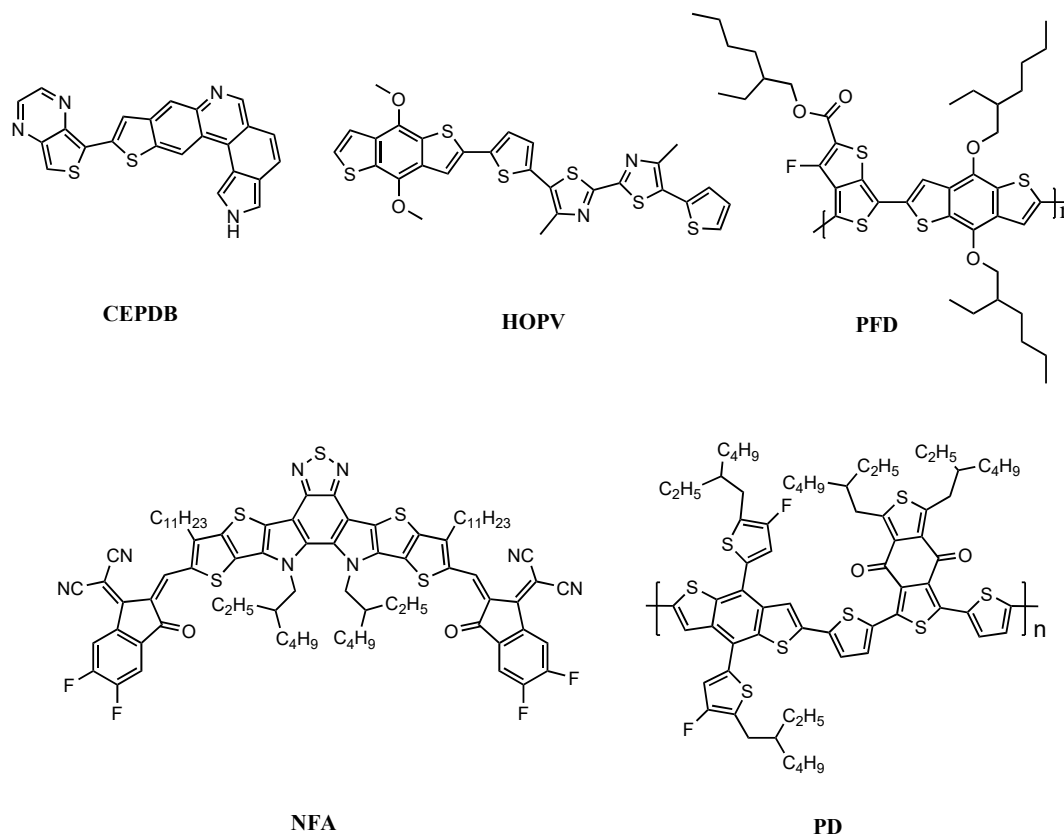


Figure 4.3: Illustration of OSC molecules sampled from the five datasets.

curated small molecules and polymers, each represented by SMILES strings, and specifically designed as p-type constituents for OSCs. The dataset provides a collection of experimentally measured properties that are critical for assessing OSC performance. Among these, the PCE is designated as the primary ground-truth property for prediction tasks, serving as the main target in subsequent evaluations.

PFD: The Polymer:Fullerene Dataset (PFD), originally reported by Nagasawa *et al.* [119], consists of 1,203 donor polymer molecules characterized by various device parameters, including PCE. These parameters were experimentally measured from older-generation polymer:fullerene-based organic solar cell (OSC) devices. Each molecule is represented by a SMILES string and assigned a unique nickname for identification. To address potential duplications caused by the same donor polymer molecule being paired with different acceptors, which can result in varying PCE values, molecules sharing the same nickname are merged. The highest PCE value for each unique molecule is selected as the ground truth. Following validation checks for SMILES string integrity using RDKit, the dataset is refined to include 1,055 valid molecules, each converted into an atom-level molecular graph using PyTorch-Geometric to facilitate graph-based modeling for accurate OSC property prediction.

NFA: The Non-Fullerene Acceptor (NFA) dataset, curated from the Polymer:NFA dataset

Table 4.1: Statistical overview of the dataset

DATASET	# GRAPHS	Avg.# Nodes	Avg.# Edges	Avg.# Rings
CEPDB	2.3M	27.6	33.3	6.7
HOPV	350	42.7	49.3	7.5
PFD	1055	77.1	84.2	8.2
NFA	654	118.2	133.0	15.8
PD	277	80.7	88.2	8.5

reported by Miyake *et al.* [115], comprises 1,318 pairs of polymer donor and non-fullerene acceptor molecules, each associated with experimentally determined PCE values. Non-fullerene acceptors represent a next-generation class of electron acceptors in organic photovoltaics, distinguished from traditional fullerene-based structures by their enhanced performance and tunability, which significantly improve the efficiency of OSC devices. During the curation process, pairs of molecules sharing the same acceptor molecule are merged based on their SMILES strings to eliminate redundancies, with the highest PCE value for each unique acceptor selected as the ground truth. Following rigorous SMILES validity checks using RDKit, the dataset is refined to include 654 valid acceptor molecules, each converted into an atom-level molecular graph using PyTorch-Geometric to enable graph-based modeling for precise prediction of OSC properties.

PD: The Polymer Donor (PD) dataset, sourced from the Polymer:NFA dataset reported by Miyake *et al.* [115], is designed for predicting OSC properties with a specific focus on polymer donor molecules, distinguishing it from the NFA dataset, which emphasizes acceptor molecules. During the curation process, pairs of molecules sharing the same donor molecule are merged based on their SMILES strings to eliminate redundancies, with the highest PCE value for each unique donor molecule selected as the ground truth. Following thorough validation of SMILES strings using RDKit, the dataset is refined to include 277 valid donor molecules, each converted into an atom-level molecular graph using PyTorch-Geometric to facilitate graph-based modeling for accurate and efficient prediction of OSC properties.

OSC Properties Details

In this work, we focus on six key properties of OSCs that are fundamental to assessing their photovoltaic behavior: power conversion efficiency, highest occupied molecular orbital, lowest unoccupied molecular orbital, band gap, open-circuit voltage, and short-circuit current density. A brief description of each property and its relevance to OSC design is given below:

- **PCE (%)** quantifies the overall efficiency of a solar cell, defined as the ratio of electrical power output to incident optical power [152]. It is regarded as the most important

performance indicator for OSCs, with values theoretically ranging between 0% and 100%.

- **HOMO** (eV) represents the highest occupied molecular orbital, while **LUMO** (eV) denotes the lowest unoccupied molecular orbital [62]. Together, they describe the frontier orbital levels of a molecule, which strongly influence its electronic and optoelectronic properties.
- **Band Gap** (eV), also known as the HOMO-LUMO gap, is defined as the energy difference between these two frontier orbitals. The gap magnitude governs optical absorption and charge transport; generally, a smaller band gap implies stronger electronic delocalization and higher reactivity [47].
- **Voc** (V) refers to the open-circuit voltage, i.e., the maximum voltage that can be extracted from a device under illumination when the current is zero [147]. It reflects the balance between photogenerated current and the built-in potential of the junction.
- **Jsc** (mA/cm²) denotes the short-circuit current density, measured when the applied voltage is zero [2]. It captures the ability of the active layer to absorb photons and generate charge carriers under illumination.

Baselines and Implementation Details. We benchmark RingFormer against eleven baseline models grouped into four categories. (i) Fingerprint-based approaches: *MACCS* [32] and *ECFP* [137]. (ii) GNN-based approaches: *GINE* [60], *GINE-VN* [43], *AttentiveFP* [183], and *O-GNN* [210]. (iii) Pooling-based approaches: *TopKPool* [38] and *SAGPool* [85]. (iv) Transformer-based approaches: *Graphormer* [189], *GPS* [133], and *GraphViT* [53]. Details of the implementation for RingFormer and all competing methods are described below.

Implementation Details of Baselines

Here we provide supplementary descriptions of the baseline models used in the experiments.

- **MACCS** belongs to the family of fingerprint-based approaches. It relies on Molecular ACCess System keys [32], which are fixed-length binary vectors (typically 166 bits). Each bit corresponds to the presence or absence of a predefined chemical fragment or structural motif within the molecule.
- **ECFP** is another fingerprinting method, based on Extended Connectivity Circular Fingerprints [137]. It encodes molecular structure by iteratively expanding neighborhoods

around each atom to identify circular substructures. The final fingerprint is a sparse binary vector in which entries denote the existence of specific atom-centered fragments.

- **GINE** is a variant of the graph isomorphism network (GIN) [184] that incorporates edge features. By embedding bond information into the message-passing scheme, GINE extends the expressive power of GIN and has shown strong performance in molecular property prediction tasks.
- **GINE-VN** [60] augments GINE with a virtual node [43], which connects to all graph nodes. This auxiliary node acts as a global memory hub, enabling each atom to both aggregate and distribute information during message passing.
- **AttentiveFP** [183] is a molecular GNN that integrates attention mechanisms. It leverages graph attention layers together with a virtual node to better capture long-range dependencies in chemical structures.
- **O-GNN** [210] is specifically designed for molecular graphs containing ring structures. In addition to updating atom and bond embeddings, O-GNN explicitly identifies rings and maintains dedicated embeddings for them, which are updated jointly during message passing.
- **TopKPool** [38] is a hierarchical pooling strategy that adaptively retains a subset of important nodes. Nodes are ranked by a trainable scoring function, and the highest-scoring ones are preserved to form a reduced graph while maintaining key structural information.
- **SAGPool** [85] is a pooling method driven by self-attention. It selects nodes using graph convolution-based attention scores, ensuring that both node features and topological context are considered when forming the pooled graph.
- **Graphormer** [189] is an early graph transformer that adapts the Transformer architecture to graph learning. It introduces node position encodings as well as distance-aware spatial encodings so that pairwise relations can be integrated directly into the attention mechanism.
- **GPS** [133] represents a recent state-of-the-art graph transformer. Each GPS layer combines a local message-passing component with a standard Transformer block, thereby integrating structural encodings with both neighborhood-level and global attention.

Table 4.2: Comparison of PCE (%) prediction performance between RingFormer and baseline methods. Results are evaluated by test MAE ($\downarrow$, where lower values indicate better performance) and reported together with the mean and standard deviation. The best outcomes are highlighted in **bold**, and the second best are underlined.

Method	CEPDB	HOPV	PFD	NFA	PD
MACCS	0.898 $\pm$ 0.001	1.632 $\pm$ 0.008	1.770$\pm$0.016	2.614 $\pm$ 0.015	2.594 $\pm$ 0.023
ECFP	0.510 $\pm$ 0.001	1.544 $\pm$ 0.026	1.787 $\pm$ 0.012	<u>2.377$\pm$0.024</u>	2.704 $\pm$ 0.016
GINE	0.460 $\pm$ 0.006	1.614 $\pm$ 0.021	1.826 $\pm$ 0.013	2.620 $\pm$ 0.035	2.528 $\pm$ 0.064
GINE-VN	0.393 $\pm$ 0.011	1.724 $\pm$ 0.020	1.878 $\pm$ 0.017	3.164 $\pm$ 0.174	2.962 $\pm$ 0.207
AttentiveFP	0.415 $\pm$ 0.009	2.002 $\pm$ 0.032	1.897 $\pm$ 0.099	2.826 $\pm$ 0.082	2.608 $\pm$ 0.029
O-GNN	0.267 $\pm$ 0.004	1.727 $\pm$ 0.073	1.868 $\pm$ 0.063	2.587 $\pm$ 0.179	2.866 $\pm$ 0.288
TopKPool	0.527 $\pm$ 0.046	1.598 $\pm$ 0.013	1.830 $\pm$ 0.008	2.644 $\pm$ 0.045	2.523 $\pm$ 0.064
SAGPool	0.536 $\pm$ 0.044	1.607 $\pm$ 0.057	1.841 $\pm$ 0.020	2.648 $\pm$ 0.101	2.557 $\pm$ 0.057
GPS	0.247 $\pm$ 0.010	1.942 $\pm$ 0.128	2.395 $\pm$ 0.221	3.233 $\pm$ 0.208	2.690 $\pm$ 0.186
Graphormer	-	1.609 $\pm$ 0.069	1.799 $\pm$ 0.047	2.689 $\pm$ 0.096	<u>2.522$\pm$0.058</u>
GraphViT	<u>0.244$\pm$0.009</u>	<u>1.479$\pm$0.061</u>	1.887 $\pm$ 0.083	2.467 $\pm$ 0.166	2.856 $\pm$ 0.089
RingFormer	0.189$\pm$0.003	1.477$\pm$0.021	<u>1.776$\pm$0.014</u>	2.259$\pm$0.012	2.482$\pm$0.047

- **GraphViT** [53] adapts vision transformer (ViT) ideas to graph learning. The graph is partitioned into smaller patches, GNN layers are applied to learn patch embeddings, and a vanilla Transformer is then used to capture long-range dependencies across patches. This design allows GraphViT to retain the strengths of both local GNNs and global Transformer attention.

4.5.2 Overall Performance

PCE prediction. We evaluate the effectiveness of RingFormer in predicting power conversion efficiency, a key performance indicator for organic solar cell molecules [154], with results reported in Table 4.2. Our experimental findings reveal that RingFormer consistently achieves superior performance across all datasets, securing the top rank on four out of five benchmarks and obtaining the runner-up position on the PFD dataset. In particular, on the large-scale CEPDB dataset, which contains millions of computationally generated OSC molecules, RingFormer achieves a test MAE of 0.189. This represents a substantial 22.8% relative improvement compared with the best-performing baseline, which attains a test MAE of 0.244, thereby underscoring RingFormer’s effectiveness in handling large-scale and highly diverse molecular spaces. Likewise, on the NFA dataset, characterized by molecules with the highest average number of

fused-ring structures (see Table 4.1), RingFormer outperforms the strong fingerprint-based baseline ECFP by 4.96%. This highlights the model’s capacity to capture complex ring connectivity and conjugation patterns that are essential for accurate prediction of photovoltaic properties.

In contrast, conventional deep learning approaches, particularly GNN-based methods such as GINE, AttentiveFP, and O-GNN, struggle to consistently outperform simple descriptor-based techniques like ECFP across the four experimentally validated datasets (HOPV, PFD, NFA, and PD). This limitation suggests that standard GNNs may fall short in modeling the hierarchical dependencies and semantic-rich substructures inherent in OSC molecules, especially when molecular architectures become larger and more complex. Fingerprint-based methods, though surprisingly competitive, lack the flexibility to generalize beyond predefined structural features and therefore exhibit limited scalability to novel chemistries.

RingFormer, by contrast, demonstrates remarkable robustness and generalizability. It not only achieves state-of-the-art results on HOPV, NFA, and PD but also maintains competitive performance on PFD, where most deep learning methods perform poorly. The consistent superiority of RingFormer across datasets of different origins, ranging from computationally derived (CEPDB) to experimentally validated (HOPV, PFD, NFA, PD), provides strong evidence of its ability to bridge the gap between in-silico modeling and real-world materials discovery. Overall, these results underscore RingFormer’s unique capability to jointly encode local atomic interactions and higher-order ring-system semantics, thereby enabling accurate, stable, and reliable prediction of PCE across diverse OSC molecular classes.

Multi-task learning. We further evaluate the performance of RingFormer within a multi-task learning framework on the CEPDB dataset, focusing on the simultaneous prediction of five key organic solar cell properties. This setting results in five regression tasks, targeting HOMO, LUMO, band gap, Voc, and Jsc. Detailed definitions and descriptions of these properties are provided in the previous section. For fair comparison, we configure the output layer of RingFormer and all deep learning baselines to match the number of target properties (five), and employ the MAE loss function for optimization.

As summarized in Table 4.3, RingFormer consistently achieves superior performance across all five regression tasks, often with substantial margins over competing methods. For instance, in predicting Jsc, RingFormer attains a test MAE of 5.993, corresponding to a 20.24% relative improvement compared with the best-performing baseline. Similarly, on HOMO and LUMO predictions, RingFormer achieves notable reductions in error, underscoring its ability to accurately capture both frontier orbital energies and device-level properties in a unified setting. These results demonstrate that RingFormer is not only effective for single-task prediction but

Table 4.3: Performance of multi-task learning on the CEPDB dataset, performance measured using test MAE ($\downarrow$; lower values reflect superior results). Units for each predicted property are listed in the second row. The best results are highlighted in **bold**, and the second best are underlined.

Method	PCE	HOMO	LUMO	Band Gap	Voc	Jsc
	%	<i>eV</i>	<i>eV</i>	<i>eV</i>	<i>V</i>	<i>mA/cm²</i>
MACCS	0.898 $\pm$ 0.001	0.115 $\pm$ 0.001	0.122 $\pm$ 0.001	0.184 $\pm$ 0.001	0.115 $\pm$ 0.001	35.297 $\pm$ 0.002
ECFP	0.510 $\pm$ 0.001	0.066 $\pm$ 0.001	0.065 $\pm$ 0.001	0.080 $\pm$ 0.001	0.066 $\pm$ 0.001	15.574 $\pm$ 0.003
GINE	0.491 $\pm$ 0.007	0.049 $\pm$ 0.001	0.058 $\pm$ 0.001	0.073 $\pm$ 0.001	0.049 $\pm$ 0.001	15.409 $\pm$ 0.292
GINE-VN	0.496 $\pm$ 0.007	0.048 $\pm$ 0.001	0.059 $\pm$ 0.003	0.073 $\pm$ 0.003	0.048 $\pm$ 0.001	15.110 $\pm$ 0.141
AttentiveFP	0.453 $\pm$ 0.018	0.041 $\pm$ 0.003	0.057 $\pm$ 0.003	0.068 $\pm$ 0.004	0.040 $\pm$ 0.003	14.182 $\pm$ 0.548
O-GNN	0.259 $\pm$ 0.008	0.026 $\pm$ 0.001	0.030 $\pm$ 0.001	0.036 $\pm$ 0.001	0.026 $\pm$ 0.001	8.039 $\pm$ 0.004
TopKPool	0.566 $\pm$ 0.047	0.053 $\pm$ 0.003	0.060 $\pm$ 0.009	0.078 $\pm$ 0.007	0.053 $\pm$ 0.003	16.120 $\pm$ 0.859
SAGPool	0.581 $\pm$ 0.009	0.056 $\pm$ 0.001	0.050 $\pm$ 0.001	0.074 $\pm$ 0.001	0.056 $\pm$ 0.001	15.510 $\pm$ 0.034
GPS	<u>0.241$\pm$0.018</u>	<u>0.020$\pm$0.003</u>	<u>0.021$\pm$0.001</u>	<u>0.025$\pm$0.001</u>	<u>0.018$\pm$0.002</u>	<u>7.514$\pm$0.350</u>
GraphViT	0.322 $\pm$ 0.050	0.040 $\pm$ 0.007	0.035 $\pm$ 0.009	0.055 $\pm$ 0.007	0.040 $\pm$ 0.007	12.479 $\pm$ 1.939
RingFormer	0.193$\pm$0.007	0.014$\pm$0.001	0.018$\pm$0.001	0.023$\pm$0.001	0.014$\pm$0.001	5.993$\pm$0.304

also robust and versatile in multi-task scenarios, where the simultaneous learning of multiple correlated properties is essential.

In addition, we observe that the GPS model consistently delivers strong results across all five properties, ranking second overall to RingFormer. This performance can be attributed to the shared design principles between GPS and RingFormer, as both combine message-passing mechanisms for modeling local atomic interactions with global attention mechanisms for capturing long-range dependencies and higher-order structural semantics, such as ring systems. The consistently superior results of RingFormer confirm the critical importance of this hybrid design, while also highlighting the additional benefits brought by its ring-level hierarchical representation and cross-attention mechanisms. Taken together, these findings validate the effectiveness of incorporating both local and global structural features in multi-task molecular property prediction. They further underscore RingFormer’s advanced capability to provide accurate, comprehensive, and generalizable predictions across multiple OSC properties, reinforcing its utility as a powerful tool for accelerating materials discovery.

Table 4.4: Ablation results for the hierarchical OSC graph, reported in terms of test MAE ($\downarrow$, lower values indicate better performance).

Variant	CEPDB	HOPV	PFD	NFA	PD
G_A only	0.550	1.912	1.841	3.122	2.683
G_R only	0.358	1.874	2.001	2.412	2.531
$G \setminus G_R$	0.606	1.526	1.831	2.304	2.540
$G \setminus G_I$	<u>0.315</u>	<u>1.497</u>	<u>1.795</u>	<u>2.299</u>	<u>2.501</u>
Full G	0.189	1.477	1.776	2.259	2.482

4.5.3 Model Analysis

Ablation on hierarchical OSC graph. We assess the effectiveness of the hierarchical OSC graph by conducting an ablation study, where the full graph representation, $G = \{G_A, G_R, G_I\}$, is compared against its individual components and reduced variants. The results are summarized in Table 4.4. We observe a clear and significant performance degradation when any of the three components is removed, thereby validating the necessity of incorporating all levels of structural information.

In particular, the variant $G \setminus G_I$, which excludes inter-level connections, still achieves superior results compared to the single-level graphs G_A and G_R . This demonstrates that modeling both atom-level and ring-level structures in tandem is essential for capturing the hierarchical semantics of OSC molecules. Furthermore, comparing the performance of the full graph G against $G \setminus G_I$ highlights the additional benefit of inter-level interactions: explicitly transferring information between atom and ring nodes further enhances predictive accuracy. Overall, these findings confirm the critical importance of the hierarchical graph design. By jointly integrating atom-level, ring-level, and inter-level structures, the proposed representation provides a more complete and chemically meaningful encoding of OSC molecules, thereby enabling more reliable and accurate property prediction.

Effectiveness of the ring-level cross-attention module. To quantify the benefit of the proposed ring-level cross-attention module, we conduct ablation studies in which it is replaced by alternative graph representation layers. Table 4.5 reports the corresponding results. In particular, the variant *Cross-attention w.o. virtual* excludes the virtual molecule node, thereby removing the pathway for efficient aggregation of global information. Meanwhile, *Vanilla Transformer* [167] denotes a standard transformer encoder applied to the graph directly, without incorporating graph-specific chemical semantics.

Table 4.5: Comparison of PCE prediction performance across alternative implementations of the ring-level graph encoder, reported in terms of test MAE ($\downarrow$, lower values indicate better performance).

Encoder	CEPDB	HOPV	PFD	NFA	PD
Cross-attention	<u>0.1886</u>	1.4774	1.7757	<u>2.2588</u>	2.4819
Cross-attention w.o. virtual	0.1860	1.5106	1.8082	2.2602	2.5876
GINE	0.3576	<u>1.4796</u>	1.7892	2.2519	2.6203
GINE-VN	0.3136	1.5069	<u>1.7856</u>	2.2665	<u>2.5296</u>
Vanilla Transformer	0.2319	1.5394	1.8273	2.4151	2.5805
GPS	0.2231	1.5137	1.7910	2.4069	2.5390

Table 4.6: Training and inference time per epoch (in seconds) across alternative ring-level graph encoder implementations.

	CEPDB		HOPV		PFD		NFA		PD	
	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train
RingFormer	71.27	504.1	0.037	0.671	0.097	1.349	0.103	0.933	0.033	0.407
RingFormer w.o. virtual	65.59	459.9	0.032	0.434	0.093	1.288	0.099	0.883	0.032	0.386
GINE	63.34	444.5	0.030	0.367	0.089	1.058	0.093	0.782	0.030	0.323
GINE-VN	65.24	440.2	0.031	0.480	0.091	1.095	0.095	0.792	0.031	0.332
Vanilla Transformer	103.5	764.5	0.048	0.639	0.109	1.704	0.116	1.179	0.042	0.507
GPS	85.71	660.1	0.048	0.687	0.118	1.948	0.123	1.353	0.047	0.599

As shown in Table 4.5, the proposed cross-attention module achieves the best performance on three out of five datasets, while also delivering competitive results on the remaining two. This consistent improvement highlights the effectiveness of explicitly incorporating edge attributes into the attention mechanism and leveraging localized neighborhood attention in place of fully connected attention. These findings confirm that the proposed design provides a stronger inductive bias for modeling the rich semantics of ring-level structures in OSC molecules.

To assess computational efficiency, we report both training and inference time per epoch (in seconds) for our method in comparison with alternative graph representation learning layers, as summarized in Table 4.6. The results indicate that the proposed ring-level cross-attention module is markedly more efficient than generic transformer-based approaches, primarily because it substitutes full pairwise attention with localized neighborhood attention supported by a virtual node. These improvements highlight that the module offers not only stronger predictive performance but also reduced computational cost, thereby establishing it as a practical and scalable

Table 4.7: Evaluation of performance when representing structures with either rings or BRICS-derived motifs, reported by test MAE ($\downarrow$, lower values indicate better performance).

	CEPDB	HOPV	PolymerFA	nNFA	pNFA
Motifs	0.4556	1.5203	1.7230	2.5013	2.5960
Rings	0.1886	1.4774	1.7757	2.2588	2.4819
Ring+Motifs	0.2440	1.4833	1.8098	2.7561	2.5908

solution for hierarchical graph modeling in OSC molecules.

Comparison to molecular motifs. To further examine whether explicitly focusing on ring systems is critical for OSC property prediction, we compare ring-specific representations with more general chemical motifs. Since rings can be regarded as a specialized subset of chemical motifs, we adapt the motif extraction approach of Zhang et al. [207], employing the BRICS algorithm [27] to decompose molecules into functional motifs. In this setting, we replace the ring-level graph G_R in RingFormer with a motif-level graph, where each node corresponds to a BRICS-derived chemical motif.

In addition to the motif-only variant, we construct a hybrid high-level graph (Ring+Motif) that integrates both ring systems and BRICS-derived motifs. In this configuration, nodes representing functional groups are connected to ring nodes if they share common atoms, thereby enabling the model to capture the interplay between aromatic cores and attached functional motifs. The results, summarized in Table 4.7, reveal that the original RingFormer, which explicitly models ring systems, consistently outperforms both the motif-only and hybrid (Ring+Motif) approaches. Although motifs contribute useful chemical semantics, the inclusion of general motifs does not provide additional benefits beyond ring-level modeling, and in some cases slightly degrades performance. This outcome underscores the central role of ring systems in determining the electronic, optical, and photovoltaic properties of OSC molecules. By prioritizing ring system patterns, RingFormer more effectively captures conjugation, electron delocalization, and structural rigidity, all of which are indispensable factors for accurate OSC property prediction.

Performance vs. ring system complexity. To further highlight RingFormer’s effectiveness in leveraging the ring structures that are intrinsic to organic solar cell molecules, we evaluate its performance across molecules with varying numbers of rings, which serve as a proxy for the complexity of their ring systems. Specifically, we examine how RingFormer’s relative improvement in MAE for PCE prediction varies across subsets of the CEPDB test set, and benchmark its performance against three representative baselines: GINE-VN, O-GNN, and GPS. As shown

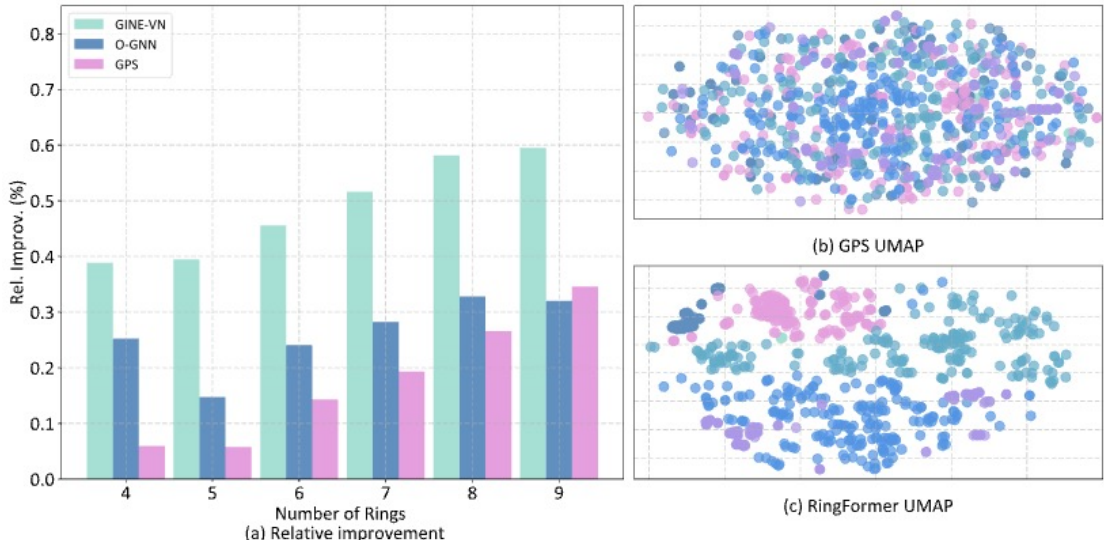


Figure 4.4: (a) Enhancement in model performance for molecules categorized by their ring counts; (b) and (c) depict embedding maps, where color encodes the number of rings.

in Figure 4.4(a), the margin of improvement achieved by RingFormer grows consistently as the number of rings in a molecule increases. This observation reveals a strong relationship between RingFormer’s predictive gains and ring system complexity, thereby confirming the necessity of explicitly incorporating hierarchical ring representations.

To further illustrate the structural patterns captured by the model, we project the graph embeddings of OSC molecules from the CEPDB test set into a low-dimensional space using UMAP [111] for dimensionality reduction. Figures 4.4(b) and (c) show that the embeddings produced by RingFormer exhibit distinct separation patterns according to the number of rings in the molecules, indicating that the model effectively encodes ring system complexity in its learned representations. By contrast, the embeddings generated by GPS display substantially less pronounced separation, suggesting that conventional transformer-based approaches struggle to capture such structural distinctions. Taken together, these results demonstrate that RingFormer not only delivers superior predictive performance as ring system complexity increases but also produces representations that are chemically meaningful, explicitly encoding the structural diversity of OSC molecules. This ability to differentiate molecules by their hierarchical ring architectures further underscores the model’s strength in capturing the conjugation, delocalization, and rigidity effects that critically influence OSC electronic and photovoltaic properties.

Performance under different GNN backbones. In the preceding experiments, we adopted GINE as the default GNN backbone for the atom-level and inter-level components of RingFormer. To further examine the robustness and versatility of the proposed framework, we conducted an

Table 4.8: Comparison of PCE prediction performance of RingFormer using alternative GNN backbones, reported in terms of test MAE ($\downarrow$, lower values indicate better performance).

	CEPDB	HOPV	PFD	NFA	PD
Best Deep Competitor	0.2442	1.4792	1.7987	2.4668	2.5221
GINE	0.1886	1.4774	1.7757	2.2588	2.4819
GatedGCN	0.1899	1.5158	1.7967	2.2742	2.4800
GraphSAGE	0.1874	1.5393	1.7923	2.3694	2.4982

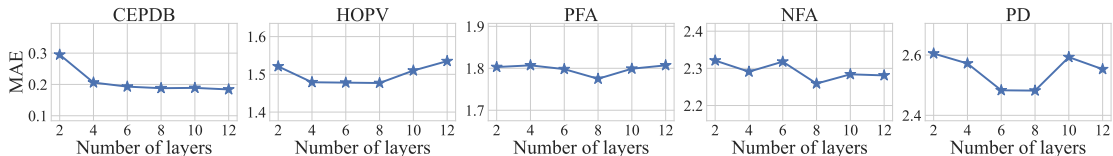


Figure 4.5: Evaluation of RingFormer’s PCE (%) predictive performance, where test MAE is reported across varying layer depths L .

additional experiment to assess its performance under different GNN backbones. The results are summarized in Table 4.8. Specifically, we implemented RingFormer with three representative GNN backbones: GINE [60], GraphSAGE [52], and GatedGCN [9]. For a comprehensive comparison, we benchmarked these variants of RingFormer against the best-performing deep learning competitor reported in Table 4.2, focusing on PCE prediction across five OSC datasets.

As shown in Table 4.8, RingFormer consistently surpasses the strongest competitor on four out of the five datasets (CEPDB, PFD, NFA, and PD), regardless of the backbone employed. This robust performance highlights the adaptability of RingFormer to different GNN architectures, demonstrating that its hierarchical design effectively leverages both local atomic interactions and higher-order ring system patterns in a backbone-agnostic manner. Among the backbones tested, RingFormer built on GINE achieves the highest performance most frequently, which can be attributed to GINE’s enhanced ability to incorporate edge attributes, thereby improving the modeling of molecular interactions that are central to OSC property prediction. Taken together, these findings emphasize that while the overall gains of RingFormer are consistent across diverse GNN backbones, choosing a backbone that captures edge-level chemical semantics (such as GINE) further strengthens the framework. This flexibility ensures that RingFormer can be seamlessly adapted to a wide range of GNN architectures, making it broadly applicable across different molecular modeling scenarios in computational materials science.

Effect of the number of layers L . To investigate the influence of model depth on predictive

performance, we systematically varied the number of RingFormer layers L from 2 to 12 and evaluated its performance on predicting PCE across the five OSC datasets (CEPDB, HOPV, PFD, NFA, and PD), using test MAE as the evaluation metric. The results are summarized in Figure 4.5.

We observe that MAE consistently decreases as L increases from 2 to 8, demonstrating that deeper architectures initially enhance predictive accuracy by enabling the model to capture more complex structural dependencies. However, when L exceeds 8, performance begins to degrade on the HOPV, PFD, NFA, and PD datasets, suggesting that excessive depth may introduce overfitting or lead to diminishing returns, particularly for smaller experimental datasets with limited data. In contrast, the CEPDB dataset exhibits a continued decrease in MAE even as L grows beyond 8. This trend can be attributed to CEPDB’s substantially larger scale (see Table 4.1 for dataset statistics), which provides sufficient training data to support deeper models without severe overfitting.

These findings underscore the importance of tailoring the number of layers to dataset size and characteristics. For large-scale datasets such as CEPDB, deeper architectures can be leveraged to extract richer hierarchical features, yielding consistent performance gains. Conversely, for smaller datasets derived from experimental measurements, moderate layer depths (around $L = 8$) strike a more favorable balance between model expressiveness and generalization, avoiding the pitfalls of over-parameterization. This analysis highlights that careful tuning of model depth is crucial for optimizing RingFormer’s performance across different OSC datasets.

Implementation Details. In RingFormer, we employ GINE [60] as the graph neural network backbone for both atom-level and inter-level components. Unless otherwise specified, the number of layers is set to $L = 8$, the hidden dimension to $d = 512$, and the number of attention heads to $C = 4$. Model training is carried out with mini-batch gradient descent using the Adam optimizer, together with a one-cycle learning rate schedule [153], where 5% of each cycle is reserved for the warm-up phase of increasing the learning rate. For the CEPDB dataset, training is performed for 30 epochs with a batch size of 1024, while for the other four datasets, we train for 100 epochs with a batch size of 32. The maximum learning rate is selected from $\{0.001, 0.0005, 0.0001, 0.00005\}$ according to validation performance. Each experiment is repeated with five random seeds, and we report the average and standard deviation of the evaluation metrics. All experiments are executed on a Linux server equipped with an Intel Xeon Gold 6226R 2.90 GHz CPU and an NVIDIA RTX 3090 GPU.

For fingerprint-based baselines (MACCS and ECFP), we generate molecular fingerprints

using DeepChem [135] and train a random forest regressor implemented in Scikit-learn [77]. Random forests have been shown in prior work [34, 115] to be strong and reliable predictors for OSC property prediction, providing a meaningful benchmark for comparison.

For deep learning baselines, including GNN-based and transformer-based architectures, we adopt publicly available implementations and harmonize their training settings with those of RingFormer to ensure a fair comparison. In particular, batch size, number of epochs, optimizer, learning rate scheduler, maximum learning rate, and initialization seeds are all aligned across methods. For GNN variants such as GINE, GINE-VN, AttentiveFP, O-GNN, TopKPool, SAGPool, Graphormer, and GPS, we fix the depth at 8 layers, the hidden dimension at 512, and the number of attention heads at 4, matching RingFormer’s configuration for comparability.

Pooling-based models such as TopKPool and SAGPool are implemented with GINE as the backbone. We apply pooling after the first four layers with a pooling ratio of 0.5, and then apply an additional four GNN layers to refine the representations. For GPS, GINE is also used as the backbone to combine local message passing with global attention. For GraphViT, we construct a hybrid architecture consisting of 4 GNN layers (GINE-based) followed by 4 transformer layers, while maintaining a hidden size of 512 and 4 attention heads. Hyperparameters for all baselines are set according to either their official implementations or the recommendations in the original publications. This uniform configuration ensures that all models are trained under comparable conditions, allowing us to draw rigorous and meaningful comparisons against RingFormer.

4.6 Conclusions

This chapter addresses the underexplored yet critical challenge of accurately predicting the device level properties of OSCs, a task that plays a pivotal role in advancing next-generation renewable energy technologies. To this end, we introduced *RingFormer*, an innovative graph transformer framework specifically designed to capture the intricate ring systems and their interconnections within OSC molecules. Unlike conventional approaches that operate primarily at the atom level, RingFormer constructs a hierarchical molecular graph representation that integrates fine-grained atomic details with higher-order ring-system architectures. This dual-level representation allows the framework to simultaneously capture local chemical interactions and global structural dependencies that are central to the electronic and optoelectronic properties of OSCs. By combining localized message-passing operations with global multi-head attention mechanisms, RingFormer learns highly expressive graph embeddings that preserve chemical fidelity while enhancing predictive capacity.

Extensive experimental evaluations across multiple OSC datasets demonstrate that RingFormer consistently surpasses state-of-the-art baselines in predicting key molecular and device-

level properties, including PCE. These results highlight the central role of ring systems in governing OSC molecular behavior and validate the necessity of hierarchical modeling for robust property prediction. Beyond benchmarking improvements, RingFormer offers practical implications for OSC research and materials discovery. Its ability to deliver accurate predictions with reduced reliance on costly and time-consuming trial-and-error synthesis underscores its potential as a powerful computational tool for guiding rational molecular design. By accelerating the identification and optimization of high-performance OSC materials, RingFormer contributes not only to the advancement of organic photovoltaics but also to the broader paradigm of AI-driven materials science. Taken together, these contributions demonstrate how hierarchical graph transformer architectures can bridge the gap between molecular complexity and predictive accuracy, providing a scalable blueprint for future computational frameworks aimed at enabling sustainable and efficient materials innovation.

Chapter 5: Hierarchical Graph Transformer with multi-level interactions

This chapter introduces HGT-MI (Hierarchical Graph Transformer with Multi-level Interactions), an AI-driven framework that significantly enhances the accuracy of predicting device-level properties for OSCs. The core innovation lies in recognizing that an OSC’s PCE depends not only on the intrinsic electronic structure of individual molecules but also on the complex, multi-scale interactions between donor and acceptor molecules. Traditional methods overlook these pairwise interactions by treating molecules in isolation, and conventional graph neural networks inherit this limitation by relying on a single, flat molecular graph. HGT-MI overcomes this bottleneck in two steps. First, we build a hierarchical OSC graph $G = \{G_A, G_M\}$, G_A is an atom-level GNN that encodes local bonding chemistry; G_M is a motif-level transformer that captures higher-order π -conjugation and side chain patterns. Then, HGT-MI extracts embeddings at each level of the hierarchical OSC graph using layer-wise message-passing mechanisms, which capture local structural patterns at the atom, motif level graphs. These embeddings are then fused into a compact molecular vector that encapsulates both local and global structural features of the OSC molecule. To model the interactions between donor and acceptor materials in OSC systems, we employ a multi-head cross-attention block that explicitly facilitates the exchange of atom-scale, motif-scale, and molecule-scale signals between the two materials. Finally, a lightweight multi-layer perceptron (MLP) regressor processes the fused representation to predict key OSC properties. HGT-MI substantially outperforms all previous methods in prediction accuracy. Furthermore, its applicability extends to other material systems, such as perovskite solar cells, light-emitting diodes, and batteries.

This chapter is based on “T. Zhang, W. Hu, Y. Sun, C. J. Zhang, Q. Li, G. Li and H. Yan, “Hierarchical Graph Transformer with Multi-level Interactions for Molecular Property Prediction in Organic Solar Cells”, *under revision*, 2025.”

5.1 Introductions

Rising demand for renewable energy prompts research into new energy resources and more efficient technologies. Solar energy is a key solution due to its cleanliness, abundance, and sustainability [24, 54, 81]. Among the promising advancements in solar technology, OSCs have garnered significant attention for their potential to be manufactured using cost-effective printing techniques on flexible substrates, making them an attractive option for scalable and versatile photovoltaic applications. Over the past decade, OSC efficiency has seen steady progress, with laboratory device efficiencies rising from approximately 10% to over 20%, driven by advancements in material design and device engineering [208, 17, 58, 200, 97, 188, 89]. Despite these improvements, a significant challenge impeding OSC development is the reliance on labor-intensive and time-consuming trial-and-error approaches for material synthesis and device fabrication [130, 160, 157, 46]. Consequently, developing robust computational methods to accurately predict material and device properties is critical to accelerating the discovery and optimization of high-efficiency OSC materials, thereby advancing their commercial viability and impact in renewable energy.

Significant advancements have been made in developing computational models for drug discovery and other biological applications, achieving notable success in the pharmaceutical industry [96, 176, 28, 67, 73]. However, OSC molecules present distinct structural characteristics that set them apart from pharmaceutical molecules, primarily due to their extended π -conjugated systems, which are critical for efficient charge transport and light absorption in photovoltaic applications [22, 45, 109, 144]. These conjugated systems, often comprising complex ring structures and delocalized electron networks, necessitate specialized modeling approaches. Consequently, traditional computational models designed for biological molecules, which typically focus on localized interactions and smaller molecular frameworks, are ill-suited for accurately predicting the properties of OSC materials or devices, underscoring the need for tailored computational strategies to address the unique structural and functional requirements of OSCs.

Among various properties of OSCs, the power conversion efficiency is the most important parameter for assessing OSC performance, representing the percentage of incident light energy converted into electrical power [166, 128, 155]. Traditionally, the PCE of OSC molecules is assessed through experiments, involving device fabrications and characterizations. This approach is not only time-consuming and costly but also requires the synthesis of a large library of OSC molecules for device evaluations. Beyond experimental approaches, current computational approaches, such as Density Functional Theory, have been employed to predict fundamental molecular properties of OSCs [69, 132, 142, 63] While DFT provides a theoretical framework for

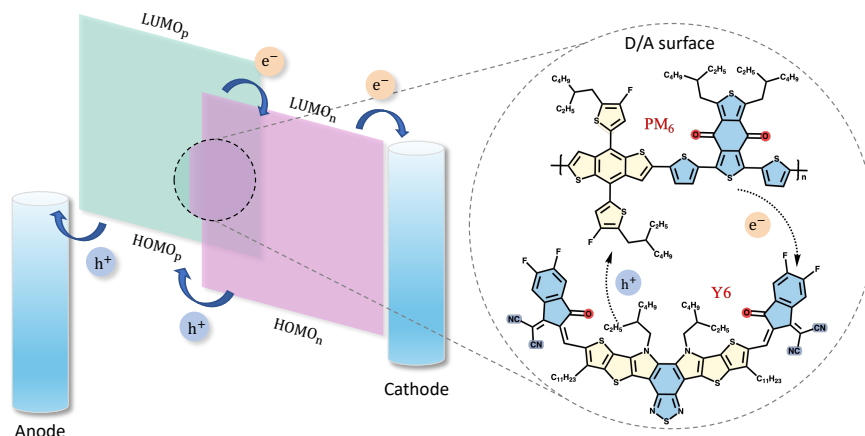


Figure 5.1: Visualization of the operational mechanism of OSC devices, showing electron-hole separation and relative energy-level alignment of donor and acceptor materials.

predicting fundamental molecular properties, it requires substantial computational resources. In addition, DFT is only suited to the property of one isolated material, while the PCE is the property of OSC devices that consist the blend of two materials, namely donor and acceptor materials [187, 106, 179]. As shown in Figure 5.1. Apparently, the interaction between the donor and acceptor is a critical factor influencing the PCE of OSC devices. However, current computational models primarily focus on individual molecular structures, often fail to capture the donor-acceptor interactions, including electronic interaction and energy-transfer dynamics between the donor and acceptor material. These interactions, critical to OSC device operation, involve electronic couplings between structural units such as aromatic rings, alkyl chains, carbonyl groups, and electron-donating or withdrawing units, all of which govern charge transfer and exciton dissociation at the molecular interface. Effective intermolecular electronic structure matching significantly influences exciton dissociation and, consequently, OSC device efficiency. Therefore, advanced modeling approaches that explicitly incorporate pairwise donor-acceptor interactions are essential for accurately representing photovoltaic behaviors and achieving reliable PCE predictions.

Recent advancements in ML and DL have introduced promising methodologies for the prediction of molecular properties [160, 155, 180, 34, 78, 1, 108, 205]. Graph neural networks, for instance, have been adopted to represent molecules as graphs, where nodes correspond to atoms and edges represent chemical bonds. However, while GNNs can capture local atomic structures, they struggle to model higher-order patterns, such as aromatic rings commonly used for OSC molecules [53]. To address this limitation, the RingFormer model, a ring-enhanced

graph transformer, was developed to explicitly incorporate structural information from aromatic ring systems, achieving superior performance compared to traditional GNNs [31]. However, RingFormer remains insufficient in modeling the critical donor-acceptor interactions in OSC devices. Other methods have specifically targeted donor-acceptor interactions. For instance, Lee et al. utilized HOMO, LUMO, and bandgap descriptors of both donors and acceptors for efficiency predictions but relied on a limited descriptor set, missing a comprehensive molecular context [87]. An interpretable ML model incorporated 240 donors, two acceptors, and ten microscopic descriptors to explicitly capture donor-acceptor interactions, yet the constrained feature space hindered its ability to fully represent complex molecular phenomena fully [1]. In other examples, Directed Message Passing Neural Networks (D-MPNNs) generate separate representations for donors and acceptors, predicting PCE through linear layers [108]. Also, D-MPNNs focus on atomic-level interactions, overlooking higher-level motif or molecular interactions between donors and acceptors. More recently, Zhang et al. compiled an experimental dataset of 1,060 realistic donor-acceptor pairs (including 672 acceptors and 240 donors) from literature, achieving a Pearson correlation coefficient (r) of 0.74 and RMSE of 2.63% for PCE prediction [206]. These efforts highlight a progression from atom-level GNNs to motif-level transformers and toward integrative D-A screening pipelines. However, these models have not fully integrated higher-order motif structures with an effective description of intermolecular interactions, hence limiting their prediction performance.

To address the aforementioned challenge, we propose HGT-MI, a novel hierarchical graph transformer framework specifically designed to enhance the representation of interactions between donor and acceptor components. HGT-MI incorporates a specialized donor-acceptor interaction mechanism that explicitly captures the relationships between these components across three hierarchical levels, atomic, motif, and molecular, enabling a comprehensive and nuanced depiction of intermolecular dynamics. By integrating these multi-level interactions, HGT-MI provides a robust representation of the structural and electronic interplay between donor and acceptor molecules, significantly improving the accuracy of OSC property predictions, such as power conversion efficiency, compared to existing computational models.

Our framework involves two primary stages:

1. **Multi-level Embedding Extraction.** A hierarchical encoder processes molecular graphs at both the atom and motif levels. Atom-level representations are generated using GNNs to capture detailed chemical environments, while motif-level representations are derived through a graph transformer to encode chemical substructure semantics. Molecular embeddings are then produced by integrating atom and motif embeddings via a multi-

head cross-attention mechanism.

2. **Multi-level Interaction Modeling.** This stage explicitly models interactions between donor and acceptor embeddings at the atom-level, motif-level, and molecular-level granularity using bi-directional multi-head cross-attention. This mechanism facilitates robust feature exchange, enhancing representations and capturing complex molecular dynamics critical for PCE predictions.

By explicitly modeling these interactions, HGT-MI significantly improves the accuracy of predicting device-level performance metrics, such as PCE. Extensive evaluations demonstrate HGT-MI’s superior predictive accuracy compared to existing computational approaches, significantly accelerating OSC material development and reducing dependence on experimental methods. Furthermore, HGT-MI’s flexible framework can extend beyond OSCs to other advanced materials such as perovskite solar cells, batteries, and LEDs.

5.2 Methodologies

5.2.1 Workflow of HGT-MI

Our HGT-MI model is constructed via two primary steps. In the first step, multi-level embeddings are extracted from the hierarchical molecular graph consisting of atom-level subgraph G_A and motif-level subgraph G_M . Local structural information in the atom-level graph is captured by GNNs, while higher-order structural motifs are extracted from the motif-level graph by graph transformers. These atom-level and motif-level embeddings are subsequently integrated into a molecular-level embedding using a multi-head cross-attention mechanism, thereby creating comprehensive hierarchical representations. In the second step, our model focuses on modeling donor-acceptor interactions at different levels (atom, motif, and molecular levels) through a multi-level interaction module. At each of these three levels, bi-directional multi-head cross-attention mechanisms alternately treat donor and acceptor embeddings as queries and keys/values, facilitating extensive and detailed information flow between donor and acceptor molecules. Finally, these interaction-enhanced embeddings are hierarchically fused, producing robust and unified representations that can significantly enhance the prediction accuracy of PCE.

5.2.2 Multi-level Embedding Extraction.

First, atom-level graphs are constructed for donor and acceptor molecules. The donor atom-level graph

$$G_{d,A} = (V_{d,A}, E_{d,A}),$$

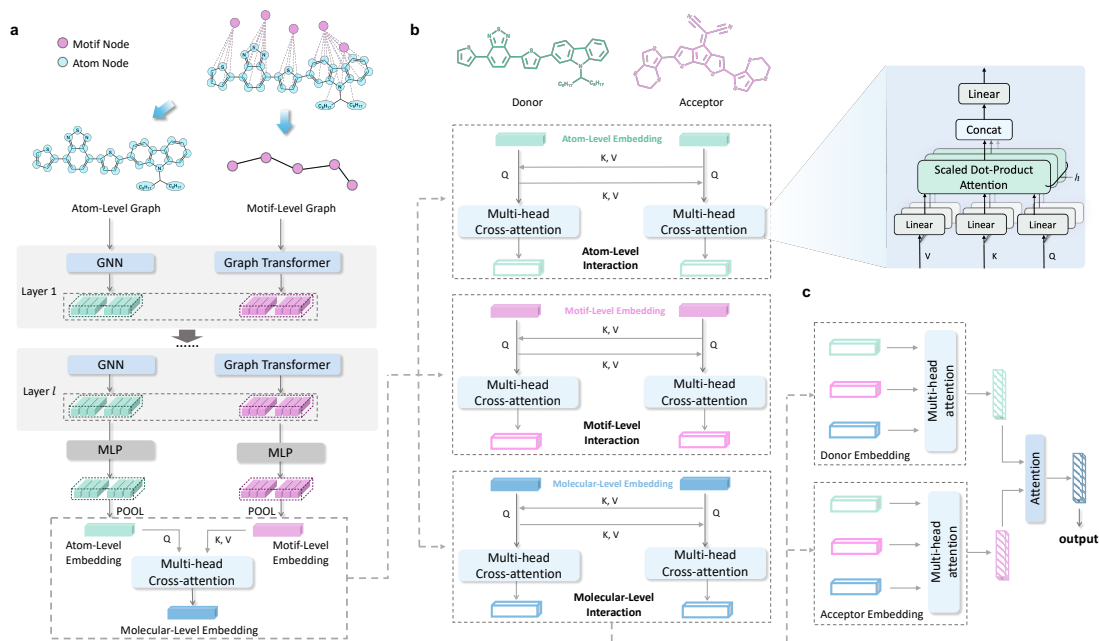


Figure 5.2: Conceptual representation of HGT-MI. (a) Multi-layer embedding derivation. (b) Cross-level interaction modeling. (c) Prediction block.

consists of atom nodes $V_{d,A}$ corresponding to individual atoms, with edges $E_{d,A}$ representing chemical bonds within the donor molecule. Similarly, the acceptor atom-level graph

$$G_{a,A} = (V_{a,A}, E_{a,A})$$

captures atom-level connectivity within the acceptor molecule. Each atom node in these graphs is described by an atom-specific feature vector, ensuring accurate representation of local chemical properties. The complete set of atom nodes is denoted

$$V_A = V_{d,A} \cup V_{a,A},$$

containing atom-level details from both donor and acceptor molecules.

Subsequently, each donor or acceptor molecule is decomposed into motif substructures using the BRICS (Breaking of Retrosynthetically Interesting Chemical Substructures) algorithm [27]. This decomposition partitions molecules into essential motifs such as aromatic rings, alkyl chains, carbonyl groups, and conjugated segments. These structural motifs are critical for property prediction because even minor changes at the motif level can significantly influence exciton dissociation and charge transport in OSCs, thereby directly affecting device performance. Illustrated in Figure 5.3 are some examples of structural motifs that have introduced significant changes in OSC performance. For instance, altering specific segments in acceptor molecules paired with a fixed donor (PM6) significantly impacts device PCE. Likewise, changing seg-

ments within donor molecules when paired with a fixed acceptor (Y6) also results in noticeable variations in PCE, underscoring the pivotal role of motif-level structural modifications.

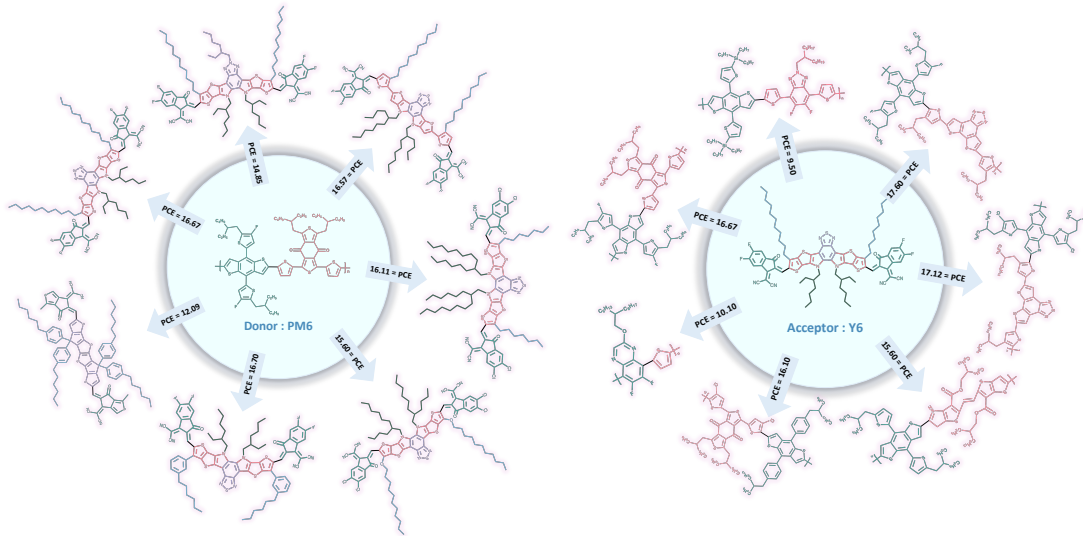


Figure 5.3: Effect of varied donor or acceptor motifs on device PCE.

Following this decomposition, motif-level graphs are constructed to capture internal motif connectivity within each molecule. The donor motif graph

$$G_{d,M} = (V_{d,M}, E_{d,M}),$$

comprises motif nodes representing substructures and edges reflecting connectivity of motifs within the donor’s chemical structure. Similarly, the acceptor motif graph

$$G_{a,M} = (V_{a,M}, E_{a,M}),$$

represents motifs and their connectivity within the acceptor molecule. Each node in these graphs is assigned a motif-specific feature vector. Together, the complete set of motif nodes is denoted as

$$V_M = V_{d,M} \cup V_{a,M},$$

encompassing motifs from both donor and acceptor molecules.

By combining atom-level graph $G_{d,A}$ and $G_{a,A}$ and motif-level graph $G_{d,M}$ and $G_{a,M}$ we construct comprehensive hierarchical OSC graphs for donor and acceptor separately:

$$G_{\text{donor}} = \{G_{d,A}, G_{d,M}\}, \quad G_{\text{acceptor}} = \{G_{a,A}, G_{a,M}\}.$$

With the hierarchical OSC graph G constructed, HGT-MI employs specialized embedding extraction layers to fully exploit the structural hierarchy for robust and comprehensive representation learning. Each layer integrates GNNs and graph transformers to capture intricate

structural patterns at both the atom and motif levels of donor and acceptor molecules in OSCs. As depicted in Figure 5.2a, each embedding extraction layer consists of two parallel modules: an atom-level GNN module, which captures fine-grained chemical environments and local bonding patterns critical for molecular properties, and a motif-level graph transformer module, which encodes intermediate structural semantics, such as ring systems and functional groups, to represent higher-order molecular features. This dual-module approach ensures that HGT-MI effectively models both localized atomic interactions and broader structural motifs, enhancing the accuracy of OSC property predictions.

These modules generate latent atom representations $\mathbf{h}_A^{(l)}$ from the GNN and motif representations $\mathbf{h}_M^{(l)}$ from the graph transformer. Following individual processing by multilayer perceptrons (MLPs), we then apply a multi-head cross-attention fusion: each pooled atom-level vector $\mathbf{z}_A$ attends to the pooled motif-level vector $\mathbf{z}_M$ to produce a refined molecular-level embedding $\mathbf{z}_{Mol}$.

Atom-Level Message Passing Module. To effectively capture local chemical structures in OSCs, we employ GNNs at the atom level [53]. Given an atom-level graph G_A . We apply GNNs to model interactions among neighboring atoms, which enables the learning of atom node representations. Specifically, in the l -th layer of our multi-level embedding extraction module, GNNs are applied as follows:

$$\mathbf{h}_{v_i,A}^{(l)} = \text{GNN}_A^{(l)}(\mathbf{h}_{v_i}^{(l-1)}, \{\mathbf{h}_{v_j}^{(l-1)} \mid v_j \in N_{G_A}(v_i)\}, \{\mathbf{e}_{ij}\}), \quad (5.1)$$

where $v_i, v_j \in V_A$ denote nodes (atoms) in G_A , $N_{G_A}(v_i)$ represents the neighbors of node v_i in G_A , $\mathbf{e}_{ij} \in E_A$ denotes the attributes of edge (bonds), and $\mathbf{h}_{v_i}^{(l-1)} \in \mathbb{R}^d$ the node representation of node v_i from the preceding layer, and d is the hidden dimension. Specifically, we have $\mathbf{h}_{v_i}^{(0)} = W_a \mathbf{x}_{v_i}$, where $W_a \in \mathbb{R}^{d \times d_{v_A}}$ are learnable parameters to transform node attributes $\mathbf{x}_{v_i}$. The function $\text{GNN}_A^{(l)}(\cdot)$ follows the Graph Isomorphism Network (GIN) framework, which we adopt as the default backbone [60]. $\mathbf{h}_{v_i,A}^{(l)}$ is the output node representation of the atom-level message passing module in the l -th layer of the multi-level embedding extraction module.

Motif-Level Message Passing Module. While the atom-level GNN captures local chemical contexts, many key optoelectronic properties of OSC molecules are governed by longer-range interactions between motifs (aromatic rings, conjugated backbones, side chains, . . .). To model these mesoscopic relationships, we construct a motif-level graph $G_M = (V_M, E_M)$, where each node $m_i \in V_M$ corresponds to a motif identified through BRICS fragmentation, and edges $e_{ij} \in E_M$ represent the inter-motif connections.

In the l -th layer of our multi-level embedding extractor, we adopt a Graph Transformer as the primary backbone at the motif level. Unlike conventional Message Passing Neural Networks (MPNNs), this layer integrates a local message-passing mechanism with a global multi-head self-attention approach [133]. This design enables each motif to interact with all others, effectively capturing both local chemical details and global structural relationships.

Local message passing. Given node features $X^{(l)} \in \mathbb{R}^{N \times d}$, edge features $E^{(l)} \in \mathbb{R}^{|E_M| \times d}$ and the adjacency $A \in \mathbb{R}^{N \times N}$, the edge-aware MPNN updates them via

$$\hat{X}_M^{(l+1)}, E^{(l+1)} = \text{MPNN}_e^l(X^{(l)}, E^{(l)}, A). \quad (5.2)$$

The provisional node embedding $\hat{X}_M^{(l+1)}$ is then combined with the residual $X^{(l)}$ and normalised:

$$X_M^{(l+1)} = \text{BN}\left(\text{Dropout}(\hat{X}_M^{(l+1)}) + X^{(l)}\right). \quad (5.3)$$

Global self-attention. Independently, a transformer block attends over all motifs:

$$\hat{X}_T^{(l+1)} = \text{GlobalAttn}^l(X^{(l)}), \quad (5.4)$$

followed by dropout, residual, and batch-norm:

$$X_T^{(l+1)} = \text{BN}\left(\text{Dropout}(\hat{X}_T^{(l+1)}) + X^{(l)}\right). \quad (5.5)$$

Fusion & MLP. The locally ($X_M^{(l+1)}$) and globally ($X_T^{(l+1)}$) updated embeddings are fused by element-wise addition and refined with a two-layer MLP:

$$X^{(l+1)} = \text{MLP}^l(X_M^{(l+1)} + X_T^{(l+1)}). \quad (5.6)$$

Because we keep $d_{l+1} = d_l = d$, the MLP only transforms the representation without changing its dimensionality.

Finally, let $\mathbf{h}_{m_i, M}^{(l)}$ denote the i -th row of $X^{(l+1)}$; this vector is the motif embedding of m_i produced by the l -th motif-level layer and serves as input to layer $l+1$.

Level-specific MLPs and pooling. After the atom- and motif-level encoders in layer l we obtain two collections of node embeddings:

$$\{\mathbf{h}_{v_i, A}^{(l)} \in \mathbb{R}^{d_l}\}_{v_i \in V_A}, \quad \{\mathbf{h}_{m_i, M}^{(l)} \in \mathbb{R}^{d_l}\}_{m_i \in V_M}, \quad (5.7)$$

where each atom $v_i \in V_A$ is encoded by the atom-level GNN on G_A and each motif $m_i \in V_M$ by the motif-level transformer on G_M .

Atom pooling. The atom representations are passed through a two-layer MLP and a global pooling operator (e.g. SUM/MEAN/SET2SET) to obtain a graph-level vector

$$\mathbf{z}_A = \text{POOL}\{\text{MLP}(\mathbf{h}_{v_i,A}^{(l)})\}_{i=1}^{N_A} \in \mathbb{R}^d. \quad (5.8)$$

Motif pooling. Analogously, motif embeddings are processed by the same architecture to yield

$$\mathbf{z}_M = \text{POOL}\{\text{MLP}(\mathbf{h}_{m_i,M}^{(l)})\}_{i=1}^{N_M} \in \mathbb{R}^d. \quad (5.9)$$

Token-splitter & cross-attention. Each d -dimensional vector $\mathbf{z}_A, \mathbf{z}_M$ is passed through a linear *token-splitter* that reshapes it into m equally sized tokens; a multi-head cross-attention module is then applied to obtain the molecular-level embedding $\mathbf{z}_{\text{Mol}}$.

5.2.3 Multi-level Interaction Modeling.

To achieve comprehensive and effective representation of donor-acceptor interaction in OSCs, we propose a hierarchical multi-level interaction mechanism that models interactions at all three levels, including atom, motif, and molecular levels. This is particularly important, as donor-acceptor interactions can be those at the atom-level (e.g., F-S interactions), or motif-level (e.g., intermolecular attraction between aromatic rings), or molecular level (e.g., miscibility of donor and acceptor molecules). As illustrated in Figure 5.2b, at each hierarchical level, we implement a bi-directional multi-head cross-attention between the two pooled embedding vectors. This structure effectively captures detailed intramolecular and intermolecular relationships, significantly improving predictive performance, especially for critical metrics like PCE.

Atom-level Interaction Modeling: Let $\mathbf{z}_{d,A}, \mathbf{z}_{a,A}$ represent the pooled atom-level embeddings for donor and acceptor molecules, respectively. We employ a bi-directional multi-head cross-attention block to model interactions. In the Donor-Acceptor direction, $\mathbf{z}_{d,A}$ acts as the query while $\mathbf{z}_{a,A}$ provides keys and values. Conversely, in the Acceptor-Donor direction, $\mathbf{z}_{a,A}$ serves as the query while $\mathbf{z}_{d,A}$ provides keys and values. This step effectively captures fine-grained chemical compatibilities such as atomic electronegativity, steric effects, and specific bonding interactions critical to donor-acceptor dynamics.

Motif-level Interaction Modeling: Motif-level interactions are modeled using pooled motif embeddings $\mathbf{z}_{d,M}$ and $\mathbf{z}_{a,M}$ derived from the graph transformer module. Utilizing the same bi-directional attention scheme, we model interactions between substructures (e.g., aromatic rings, alkyl side chains, carbonyl groups). Each donor motif embedding queries acceptor motif embeddings and vice versa, explicitly modeling how higher-order structural motifs interact.

Molecular-level Interaction Modeling: Finally, molecular-level donor-acceptor interactions are modeled by directly cross-attending molecular-level embeddings. Specifically, after obtaining molecular embeddings $\mathbf{z}_{d,Mol}$, $\mathbf{z}_{a,Mol}$ through multi-head fusion, we apply a bi-directional multi-head attention block. Here, the donor molecular embedding queries the acceptor molecular embedding and vice versa, capturing comprehensive global interactions relevant to device performance.

Interaction module. Let $\mathbf{z}_{d,A}$, $\mathbf{z}_{a,A}$ denote the pooled atom-level representations for donor and acceptor molecules, respectively; $\mathbf{z}_{d,M}$, $\mathbf{z}_{a,M}$ the pooled motif-level vectors; and $\mathbf{z}_{d,Mol}$, $\mathbf{z}_{a,Mol}$ the molecular-level representations. We apply multi-head cross-attention between donor and acceptor representations at each hierarchical level.

Atom-Level Cross-Attention: for head $c = 1, \dots, h$, compute

$$\begin{aligned} \mathbf{Q}_{d,A,c} &= \mathbf{W}_c^Q \mathbf{z}_{d,A}, & \mathbf{K}_{a,A,c} &= \mathbf{W}_c^K \mathbf{z}_{a,A}, & \mathbf{V}_{a,A,c} &= \mathbf{W}_c^V \mathbf{z}_{a,A}, \\ \alpha_{d,A,c} &= \text{Softmax}\left(\frac{\mathbf{Q}_{d,A,c} \mathbf{K}_{a,A,c}^\top}{\sqrt{d_{\text{head}}}}\right), & \mathbf{head}_{d,A,c} &= \alpha_{d,A,c} \mathbf{V}_{a,A,c}. \end{aligned} \quad (5.10)$$

Concatenate all heads:

$$\tilde{\mathbf{z}}_{d,A} = [\mathbf{head}_{d,A,1} \parallel \dots \parallel \mathbf{head}_{d,A,h}], \quad \mathbf{z}'_{d,A} = \mathbf{W}_A^O \tilde{\mathbf{z}}_{d,A}. \quad (5.11)$$

Acceptor-as-query (symmetric) step:

$$\begin{aligned} \mathbf{Q}_{a,A,c} &= \mathbf{W}_c^Q \mathbf{z}_{a,A}, & \mathbf{K}_{d,A,c} &= \mathbf{W}_c^K \mathbf{z}_{d,A}, & \mathbf{V}_{d,A,c} &= \mathbf{W}_c^V \mathbf{z}_{d,A}, \\ \alpha_{a,A,c} &= \text{Softmax}\left(\frac{\mathbf{Q}_{a,A,c} \mathbf{K}_{d,A,c}^\top}{\sqrt{d_{\text{head}}}}\right), & \mathbf{head}_{a,A,c} &= \alpha_{a,A,c} \mathbf{V}_{d,A,c}, \end{aligned} \quad (5.12)$$

Concatenate all heads:

$$\tilde{\mathbf{z}}_{a,A} = [\mathbf{head}_{a,A,1} \parallel \dots \parallel \mathbf{head}_{a,A,h}], \quad \mathbf{z}'_{a,A} = \mathbf{W}_A^O \tilde{\mathbf{z}}_{a,A}. \quad (5.13)$$

Motif-Level Cross-Attention: Analogous computations are conducted at the motif-level, replace $\mathbf{z}_{d,A}$, $\mathbf{z}_{a,A}$ with motif-level vectors $\mathbf{z}_{d,M}$, $\mathbf{z}_{a,M}$.

Molecular-Level Cross-Attention: Similar computations are performed using molecular-level summaries $\mathbf{z}_{d,Mol}$, $\mathbf{z}_{a,Mol}$.

In each case, the weight matrices $\mathbf{W}_c^Q$, $\mathbf{W}_c^K$, $\mathbf{W}_c^V$ and output projections $\mathbf{W}_{(\cdot)}^O$ are learned separately at each hierarchical level (atom, motif, and molecular).

5.2.4 Hierarchical Fusion and Prediction

Following the generation of refined donor and acceptor embeddings that capture intermolecular interactions at the atom, motif, and molecular levels, these embeddings are integrated to form

unified molecular representations for OSC property prediction. For each molecule (donor and acceptor independently), the three hierarchical embeddings, atom-level, motif-level, and molecular-level, are stacked to create a comprehensive multi-scale representation. These stacked embeddings are then processed through a multi-head attention mechanism, which effectively weighs and combines the structural information across different hierarchical levels, producing robust and expressive donor and acceptor representations. Subsequently, an additional attention mechanism integrates these final donor and acceptor embeddings, yielding a joint molecular representation $\mathbf{h}_G$ that encapsulates the combined structural and electronic features of both molecules. This joint representation serves as a holistic descriptor of the donor-acceptor system, enabling accurate prediction of OSC properties such as power conversion efficiency.

Given that the OSC target properties are continuous, we frame the prediction as a regression task. The unified embedding $\mathbf{h}_G$ is projected through a linear layer to predict the target property $\hat{y}_G$. The model is optimized during training by minimizing the MAE:

$$L_{\text{MAE}} = \frac{1}{B} \sum_{b=1}^B |y_{G_b} - \hat{y}_{G_b}|, \quad (5.14)$$

where B denotes the batch size, and y_{G_b} and $\hat{y}_{G_b}$ represent the ground-truth and predicted values for the b -th donor-acceptor molecular pair, respectively.

5.3 Results

To thoroughly evaluate the predictive capabilities of HGT-MI for OSC property prediction, we employ two distinct datasets of donor-acceptor pairs, as summarized in Table 5.1. The first dataset comprises 51,000 D-A pairs compiled from existing literature [103], providing a large-scale, computationally derived benchmark for assessing model performance. The second dataset, an experimentally validated set of 1,200 D-A pairs, was manually curated and verified by our team using data from peer-reviewed publications, ensuring high-quality, real-world relevance. Additional details on the construction and preprocessing of these datasets, including data cleaning, SMILES string validation, and graph conversion processes, are provided below to clarify the methodology and ensure reproducibility of the evaluation.

- DA Pair 51K:** This dataset originates from a large-scale virtual screening study by Lopez *et al.* [103]. It contains over 51000 D-A pairs, in each pair, the donor molecule is fixed as PCDTBT, a well-characterized polymer, while the acceptor is a unique non-fullerene acceptor. The NFAs were systematically generated through combinatorial assembly of 107 common molecular fragments, including core units, spacers, and terminal groups, and are represented as SMILES strings. Molecular structures were validated using RDKit to

Table 5.1: Dataset statistics.

Dataset	# Molecules	Avg. #Nodes	Avg. #Edges	#Motifs
DA_pair_51K	51K $\times$ 2	85	191.7	11.8
DA_pair_1.2K	1.2k $\times$ 2	191	424	23.6

ensure chemical validity. Key electronic properties, including HOMO, LUMO, and band gap, were computed using DFT methods. PCE for each D-A pair was estimated using a modified Scharber model, which provides a theoretical approximation of photovoltaic performance. During preprocessing, samples with negative or anomalously low PCE values were removed to ensure data quality, resulting in a refined dataset well-suited for modeling and predicting OSC performance.

- DA_Pair_1.2k:** This dataset was meticulously curated from the OSC literature to provide a high-quality, experimentally validated benchmark for OSC research. The initial collection included 1,700 experimentally reported D-A pairs, sourced from peer-reviewed publications. During preprocessing, duplicate entries and samples with negative or extremely low PCE values were removed to ensure data reliability and relevance. The resulting dataset comprises 1,200 unique D-A pairs, each represented by two SMILES strings corresponding to the donor and acceptor molecules, respectively. This dataset serves as a realistic and robust benchmark for evaluating computational models, such as HGT-MI, that aim to capture the complex intermolecular interactions between donor and acceptor molecules critical for accurate prediction of OSC performance metrics like PCE.

Table 5.2 presents a comprehensive evaluation of our proposed HGT-MI framework on two D-A datasets, DA_Pair_51K and DA_Pair_1.2K, benchmarking its performance against several representative baseline models, including MACCS [32], GINE [60], AttentiveFP [183], TopKPool [38], GPS [133], Graphormer [189], and RingFormer [31]. On the large-scale DA_Pair_51K dataset, HGT-MI achieves the best predictive performance, recording the lowest MAE of 0.334 and MSE of 0.379. These results reflect significant improvements over other baseline methods. Specifically, HGT-MI outperforms the second-best model, RingFormer, by achieving a 15.2% lower MAE and a 16.3% lower MSE, demonstrating its superior ability to model complex molecular interactions. On the smaller, manually curated DA_Pair_1.2K dataset, HGT-MI continues to exhibit exceptional performance, with an MAE of 1.576 and an MSE of 3.975. Compared to the next-best performing model, Graphormer (MAE 1.961, MSE 6.476), HGT-MI achieves a 19.6% reduction in MAE and a 38.6% reduction in MSE. Furthermore,

Table 5.2: PCE prediction performance comparison of HGT-MI and baseline methods on D-A datasets. Metrics are MAE ($\downarrow$) and MSE ($\downarrow$); values are mean $\pm$ standard deviation. $\downarrow$ indicates smaller values are better.

Methods	DA_Pair_51K		DA_Pair_1.2K	
	MAE($\downarrow$)	MSE($\downarrow$)	MAE($\downarrow$)	MSE($\downarrow$)
MACCS	0.669 $\pm$ 0.004	1.162 $\pm$ 0.016	2.103 $\pm$ 0.003	6.932 $\pm$ 0.026
GINE	0.677 $\pm$ 0.008	1.227 $\pm$ 0.042	2.085 $\pm$ 0.053	7.616 $\pm$ 0.556
AttentiveFP	0.931 $\pm$ 0.029	2.165 $\pm$ 0.206	3.996 $\pm$ 0.166	19.023 $\pm$ 3.972
TopKPool	0.766 $\pm$ 0.009	1.470 $\pm$ 0.058	4.358 $\pm$ 1.385	21.514 $\pm$ 4.657
GPS	0.564 $\pm$ 0.024	0.776 $\pm$ 0.088	3.150 $\pm$ 0.627	14.838 $\pm$ 4.186
Graphormer	0.519 $\pm$ 0.007	0.791 $\pm$ 0.016	<u>1.961</u> $\pm$ 0.044	<u>6.476</u> $\pm$ 0.553
RingFormer	<u>0.394</u> $\pm$ 0.024	<u>0.453</u> $\pm$ 0.040	2.017 $\pm$ 0.029	7.256 $\pm$ 0.346
HGT-MI (ours)	0.334 $\pm$ 0.009	0.379 $\pm$ 0.018	1.576 $\pm$ 0.065	3.975 $\pm$ 0.269

HGT-MI surpasses RingFormer (MAE 2.017, MSE 7.256) by 21.9% in MAE and 45.2% in MSE, underscoring its robustness even on smaller, experimentally validated datasets.

The consistently superior predictive performance of HGT-MI across both the large-scale DA_Pair_51K and the high-quality DA_Pair_1.2K datasets highlights the effectiveness of its hierarchical, multi-level interaction architecture. By explicitly modeling intra-molecular motifs and intricate intermolecular donor-acceptor interactions at the atom, motif, and molecular levels, HGT-MI significantly enhances the accuracy of PCE predictions. These results demonstrate HGT-MI’s robust consistency, scalability, and ability to outperform state-of-the-art graph-based approaches, making it a powerful tool for accelerating the discovery and optimization of organic solar cell materials.

To elucidate the relative contributions of the atom-level, motif-level, and molecular-level feature extraction components in HGT-MI, we conducted ablation studies by systematically removing each level and assessing the resulting impact on predictive performance for OSC property prediction. As illustrated in Figure 5.4, the results reveal a significant performance degradation when any submodule of the hierarchical graph G is excluded, underscoring the necessity of integrating all three levels for optimal accuracy. Notably, the removal of the motif-level module ($G \setminus G_M$) leads to the most pronounced decline in performance, with errors increasing to 0.874 (MAE) and 1.742 (MSE) on the DA_Pair_51K dataset, and 2.142 (MAE) and 7.702 (MSE) on the DA_Pair_1.2K dataset. This substantial drop highlights the critical

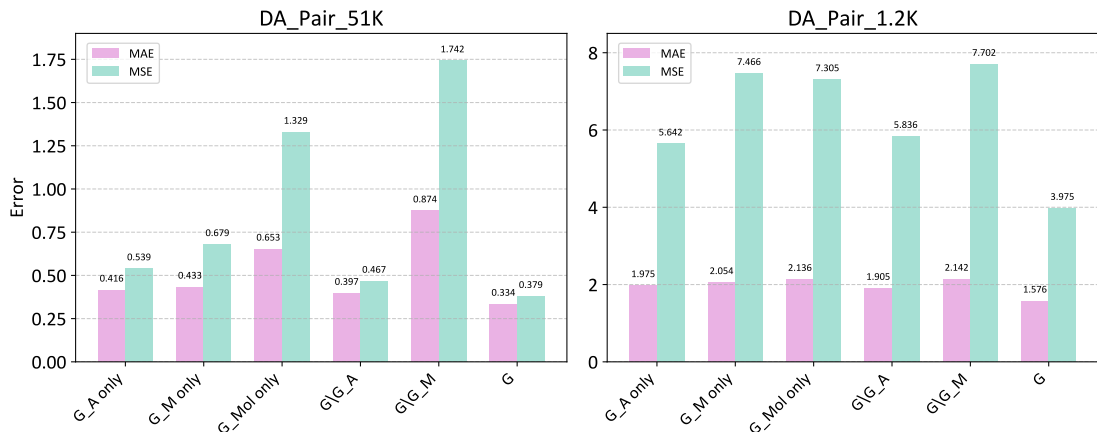


Figure 5.4: Effect of removing individual components on hierarchical OSC graph performance.

role of motif-level representations in capturing intermediate structural semantics, such as ring systems and functional groups, which are pivotal for OSC properties like PCE. Single-level variants, which rely solely on atom, motif, or molecular features, consistently exhibit inferior performance compared to the full model, confirming that no single scale is sufficient to capture the complexity of OSC molecules. These findings emphasize the synergistic importance of integrating atom-level, motif-level, and molecular-level contexts to achieve precise and robust PCE predictions in OSCs.

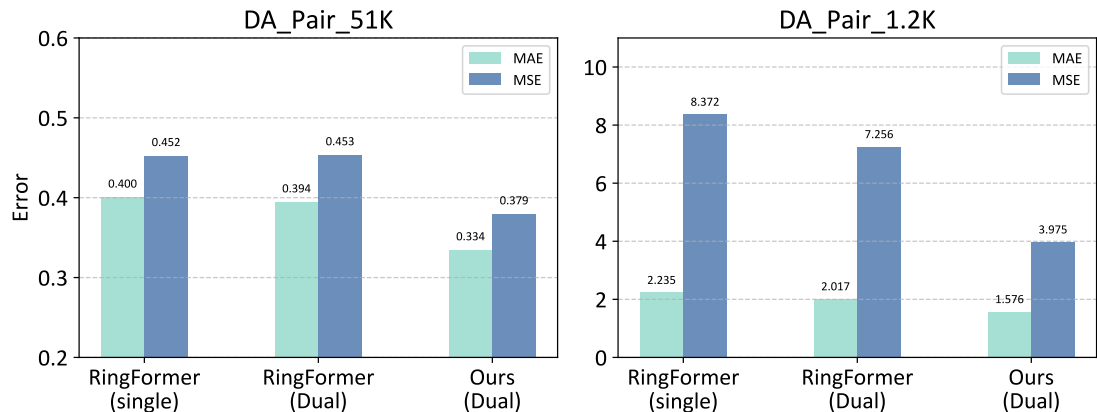


Figure 5.5: Performance comparison between single-molecule and donor-acceptor representations.

To investigate the critical role of donor-acceptor interactions in predicting device performance for OSCs, we compare the predictive performance of our HGT-MI model, which explicitly emphasizes D-A interactions, against simplified models that rely solely on single-molecule representations. For instance, the previously reported RingFormer model [31] was designed to predict

OSC molecular properties using individual molecule representations (RingFormer-single). As illustrated in Figure 5.5, such models inherently fail to capture essential intermolecular interactions between donor and acceptor molecules, leading to significantly higher prediction errors for device-level properties like PCE. While a dual-molecule variant of RingFormer (RingFormer-dual) improves performance by incorporating combined D-A representations, our proposed HGT-MI framework achieves even greater accuracy through its hierarchical, multi-level interaction modeling. Specifically, on the DA_Pair_51K dataset, HGT-MI reduces the MAE by approximately 15% and the MSE by 16% compared to RingFormer-dual. On the smaller, experimentally validated DA_Pair_1.2K dataset, the improvements are more substantial, with HGT-MI achieving approximately 22% lower MAE and 45% lower MSE. These results underscore the necessity and effectiveness of explicitly modeling fine-grained motif-level interactions and hierarchical relationships across atom, motif, and molecular levels between donor and acceptor pairs, enabling HGT-MI to deliver highly accurate PCE predictions for OSC materials.

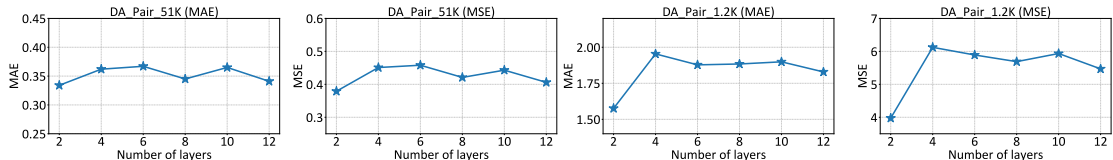


Figure 5.6: PCE prediction performance of HGT-MI, shown as test MAE and MSE, for varying numbers of layers (l)

Finally, we evaluated the impact of layer depth l on the performance of our HGT-MI model for predicting PCE in OSCs, we conducted experiments using the DA_Pair_51K and DA_Pair_1.2K datasets, with results illustrated in Figure 5.6. On the larger DA_Pair_51K dataset, the shallowest model ($l = 2$) achieves the lowest MAE of 0.334, indicating strong performance with minimal layers. Increasing the number of layers to 4 and 6 results in higher MAEs of 0.362 and 0.367, respectively, with a slight improvement at 8 layers (MAE 0.345), followed by fluctuations at 10 layers (MAE 0.365) and 12 layers (MAE 0.341). On the smaller, experimentally validated DA_Pair_1.2K dataset, the MAE rises from 1.576 at ($l = 2$) to 1.954 at ($l = 4$), remains elevated through ($l = 10$), and shows a modest improvement at 12 layers (MAE 1.829). These results suggest that deeper architectures do not consistently enhance predictive accuracy and may lead to performance degradation, particularly on smaller datasets, potentially due to overfitting or increased model complexity. Consequently, a carefully tuned network depth, typically favoring shallower models around ($l = 2$), is crucial for effectively capturing atom-level, motif-level, and molecular-level interactions to achieve optimal PCE predictions in OSCs.

5.4 Discussion

Predicting device-level properties of OSCs, such as PCE, requires more than modeling the structural features of isolated molecules. While single-molecule models effectively capture intrinsic electronic properties, such as frontier orbital energies (HOMO and LUMO) and molecular band gaps influenced by ring conjugation, substituents, and molecular conformations, they inherently fail to account for critical intermolecular interactions that are pivotal in OSC systems comprising donor and acceptor material components. Specifically, the efficiency of an OSC device is profoundly influenced by photophysical processes at the donor-acceptor interface, including charge separation, exciton diffusion, and charge transport. These processes are modulated by intermolecular interactions such as π - π stacking, hydrogen bonding, electrostatic complementarity, and the relative molecular orientations within donor-acceptor blends. Such interactions, which significantly impact device performance, are inaccessible when only single-molecule features are considered.

Moreover, capturing donor-acceptor interactions at a single resolution or granularity is insufficient to fully represent the complex interplay of chemical phenomena occurring across multiple scales. At the atom level, subtle electronic couplings between frontier orbitals drive initial charge transfer processes, laying the foundation for efficient charge separation. At the motif level, intermediate-range interactions, such as non-covalent interactions involving ring systems or functional groups, shape film morphology and micro-phase separation, directly affecting exciton dissociation and charge transport pathways. At the molecular level, large-scale properties, including relative molecular orientations and dipole interactions, define the bulk heterojunction architecture, critically influencing overall device performance.

To address these challenges, our proposed HGT-MI framework explicitly integrates these hierarchical interactions through multi-level representations spanning atom-level, motif-level, and molecular-level features. By employing a bi-directional, multi-head cross-attention mechanism that operates across these granularities, HGT-MI ensures a comprehensive and dynamic flow of information between donor and acceptor molecules. This approach effectively captures the intricate chemical synergies that govern OSC device efficiency, leading to significantly improved predictive accuracy. Extensive evaluations demonstrate that HGT-MI outperforms single-molecule models and other baseline approaches, providing deeper insights into molecular design strategies for developing high-performance OSC materials and advancing their optimization for enhanced photovoltaic applications.

5.5 Conclusions

We introduce HGT-MI, a hierarchical graph transformer with a multi-level interactions framework specifically designed to improve the prediction accuracy of OSC properties by effectively integrating donor-acceptor interactions at different levels. In the first stage, multi-level embedding extraction, each molecule is encoded at the atom-level via a GNN and at the motif-level via a graph transformer; these representations are fused by cross-attention into a context-rich molecular-level embedding. In the second stage, multi-level interaction modeling, donor and acceptor embeddings interact via bi-directional multi-head cross-attention at the atom, motif, and molecular scales. Extensive experiments on two D-A pair datasets demonstrate that HGT-MI achieves superior accuracy in PCE prediction compared to existing baselines. By integrating hierarchical intramolecular cues with explicit intermolecular interaction modeling, HGT-MI offers accurate prediction capability that would greatly accelerate the screening and discovery of high-performance OSC materials, speeding up the development of next-generation organic photovoltaics. Furthermore, due to their flexible, hierarchical design, our approaches can be adapted across a wide spectrum of advanced materials, ranging from perovskite solar cells to next-generation battery chemistries.

Chapter 6: Discussion and conclusion

The field of molecular property prediction has its origins in the initial stages of computational chemistry and drug discovery, where early efforts focused on developing methods to forecast the behavior and characteristics of small molecules for pharmaceutical applications. In earlier periods, the advent of quantitative structure-activity relationship models marked a foundational step, relying on simple linear regressions to correlate molecular structures with biological activities, primarily for drug candidates. These models, pioneered by researchers like Hansch and Fujita in earlier times, used basic descriptors such as lipophilicity and electronic parameters to predict drug potency and toxicity. As computational power grew in the 1980s and 1990s, the integration of more sophisticated techniques, including 3D-QSAR and comparative molecular field analysis, allowed for better handling of spatial conformations and steric effects, enabling predictions of binding affinities and pharmacokinetic properties. The turn of the millennium saw a shift towards ML-driven approaches, with support vector machines and random forests becoming popular for handling larger datasets from high-throughput screening. By the 2010s, deep learning revolutionized the domain, with convolutional neural networks and recurrent neural networks applied to SMILES strings and molecular graphs for tasks like solubility, toxicity, and ADME (absorption, distribution, metabolism, excretion) prediction. Recent advancements have emphasized graph neural networks and transformers, as seen in models like PotentialNet and ImageMol, which leverage pretraining on massive unlabeled datasets to achieve state-of-the-art accuracy in drug-target interaction and property prediction. These developments have accelerated drug discovery by reducing the need for costly wet-lab experiments, with applications in virtual screening and lead optimization, though challenges like data scarcity and model interpretability persist.

While molecular property prediction has matured significantly in drug discovery, its application to materials science, particularly for OSCs, faces distinct challenges and opportunities due to the unique structural and functional requirements of materials. Unlike pharmaceutical molecules, which primarily interact with biological systems, OSC materials demand predictions of physical, electronic, and optoelectronic properties that govern device performance under diverse operational conditions, such as light absorption, charge transport, and long-term stability. These properties are heavily influenced by complex molecular architectures, including extended π -conjugated systems, intermolecular donor-acceptor interactions, and hierarchical structural motifs like ring systems and functional groups. The development of materials science, particularly for renewable energy applications like OSCs, has historically relied on experimental synthesis and characterization, which are time-consuming and costly. Over the past few decades,

advancements in computational methods have begun to transform this landscape, with significant progress in modeling material properties to accelerate discovery and optimization. The integration of artificial intelligence into materials science, often referred to as “AI for Materials”, has emerged as a powerful paradigm to address these challenges, leveraging AI-driven approaches to predict material properties with unprecedented accuracy and efficiency.

Evolution of Materials Development. Since the establishment of the bulk-heterojunction architecture in 1995, organic solar cells have advanced through distinct material eras, beginning with polymer-fullerene blends that enabled efficient exciton dissociation at distributed donor-acceptor interfaces. Early demonstrations (e.g., MEH-PPV:PC₆₁BM) cemented BHJ as the default active-layer paradigm and framed key device-physics considerations for the following decades. As performance matured, fullerene systems revealed intrinsic limitations, large voltage losses, weak and narrow absorption, and limited tunability, prompting a pivot toward molecularly designable, non-fullerene small-molecule acceptors (SMAs).

The SMA era opened with ITIC-type A-D-A frameworks and accelerated with the Y6 family of low-bandgap fused-ring acceptors. Relative to ITIC, Y6 derivatives delivered systematically smaller voltage losses through reduced nonradiative recombination, while sustaining strong absorption and favorable energy-level alignment. Design rules diversified rapidly, tuning the central core, end groups, and side-chain architecture, to co-optimize packing, crystallinity, and film processability; the L8-BO variant, for example, tightened π - π packing and elevated fill factor when paired with D18, helping polymer:SMA devices surpass around 19% under laboratory conditions. In parallel with polymer:SMA systems, the materials landscape broadened into all-polymer and all-small-molecule OSCs as well as multicomponent “organic alloy” blends. All-polymer cells offer standout thermal and photostability, often retaining a high fraction of initial efficiency after extended high-temperature operation, though charge-transport and morphology control remain challenges; all-small-molecule systems deliver batch reproducibility but can face processing and stability hurdles; and multicomponent blends widen spectral coverage and enable compositional tuning of energy levels and crystallinity at the cost of increased morphological complexity.

Concurrently, application spaces have clarified practical targets for materials development. OSCs combine low embodied energy, mechanical flexibility, and lightweight form factors without toxic elements, traits attractive for urban deployment and scenarios where conventional silicon is impractical. Indoor organic solar cells have emerged as a particularly promising near-term market: by tailoring absorption to indoor spectra, OSCs routinely exceed 30% under indoor lighting (versus around 6% for silicon), with stability requirements that are more compatible with

current organic materials. Despite these advances, lab-to-fab translation still hinges on solvent systems and processing windows, device stacks (e.g., inverted architectures), and interlayer compatibility; moreover, some of the highest-efficiency polymer:SMA systems continue to face stability and manufacturability constraints highlighted across category-wise comparisons.

In summary, the evolution from polymer-fullerene to Y6-class SMAs, and onward to all-polymer, all-small-molecule, and multicomponent architectures has reframed OSC materials development as a coordinated exercise in backbone design, side-chain engineering, and morphology control. Going forward, the central levers for progress are clear: molecular tunability to minimize voltage loss, blend strategies to manage morphology across processing conditions, and application-specific targeting (especially indoor OSCs) that balances efficiency, stability, and scalable manufacturing.

AI for Materials. The emergence of AI for materials has revolutionized the approach to these challenges, offering AI-driven methods to predict material properties with greater accuracy and scalability. Unlike traditional DFT, which relies on quantum mechanical calculations, AI-based models leverage large datasets to learn complex relationships between molecular structures and their properties. Early applications of AI in materials science adapted techniques from drug discovery, such as QSAR models and molecular fingerprints (e.g., ECFP, MACCS), to predict basic properties like solubility or band gap. However, these methods struggled to capture the structural complexities of OSC materials, particularly the extended conjugated systems and intermolecular interactions unique to photovoltaic applications. The advent of graph neural networks marked a significant advancement, as they model molecules as graphs with atoms as nodes and bonds as edges, naturally representing chemical structures. GNN-based models like GINE and AttentiveFP have shown promise in predicting OSC properties by capturing local atomic environments and short-range interactions. More recently, transformer-based architectures, such as Graphormer and GPS, have incorporated global attention mechanisms to model long-range dependencies, improving predictions for properties influenced by molecular topology, such as PCE. These models have achieved notable success, with some studies reporting MAE reductions of 10-15% compared to DFT for energy-level predictions on OSC datasets. Additionally, hybrid ML-DFT frameworks have emerged to combine the precision of quantum calculations with the speed of ML, particularly for indoor OSC applications, where specific light absorption profiles are critical. A key trend in AI for materials is the development of hierarchical and multi-scale models to address the limitations of single-resolution approaches. For instance, descriptors like Ring2Vec have been proposed to capture both atom-level and motif-level features, such as ring systems, which are critical for OSC performance due to their role in π -conjugation. To motivate

the emphasis on ring-level representations, it is necessary to clarify their physical meaning and relevance in organic semiconductors.

Ring structures, particularly aromatic and fused-ring motifs, play a central role in determining the electronic and optoelectronic properties of organic semiconductor materials. In organic solar cells, these ring-based π -conjugated units govern key processes such as charge delocalization, energy level alignment, and optical absorption. As a result, the arrangement and interaction of ring structures strongly influence charge generation, transport, and recombination in donor-acceptor systems. Compared to atom-level representations, ring-level motifs provide a more chemically meaningful abstraction of molecular structure. Rings often correspond to stable functional subunits that are preserved across different molecular designs and are commonly targeted in experimental molecular engineering. By modeling molecules at the ring level, the proposed framework captures long-range conjugation patterns and structural motifs that are difficult to infer from isolated atoms or bonds alone. Furthermore, explicitly incorporating ring-level representations enhances the model’s ability to capture interaction patterns between donor and acceptor molecules. Inter-ring interactions, such as π - π stacking and conjugation alignment, are critical determinants of charge-transfer efficiency in organic solar cells. By learning representations that reflect these motif-level interactions, the model achieves improved predictive accuracy and offers better interpretability, as the learned features can be more directly associated with chemically and physically meaningful structural elements. Importantly, this ring-level modeling strategy aligns well with experimental intuition in organic electronics, where performance optimization typically focuses on modifying conjugated cores or fused-ring backbones rather than individual atoms. Therefore, explicitly modeling ring structures provides a principled foundation for the ring-based representations and hierarchical architectures developed in this thesis.

Similarly, explainable AI approaches have been explored to provide insights into how molecular features contribute to PCE, enabling researchers to design molecules with targeted properties. Pretraining on large, diverse datasets, such as those from the Clean Energy Project, has further enhanced model generalizability, allowing predictions to generalize across different OSC material classes, including polymers and small molecules. Despite these advances, challenges remain, including the lack of interpretability in many AI models, the need for high-quality experimental data, and the difficulty of modeling intermolecular D-A interactions, which are essential for device-level property predictions. These gaps highlight the need for innovative frameworks that integrate multi-scale representations and explicitly address intermolecular dynamics, setting the stage for the contributions of this thesis.

Recent advancements in predicting material properties for OSCs have harnessed machine

learning to address longstanding challenges. Several studies have employed supervised learning methods, such as support vector machines, to categorize OSC materials into high- and low-performance groups based on molecular descriptors, facilitating predictions of whether novel donor/acceptor pairs will achieve high power conversion efficiency. One notable approach utilized SVMs to identify molecular structures likely to yield high-efficiency donor or acceptor materials, thereby reducing dependence on extensive experimental trials. Random forests and gradient boosting regression trees have been applied to predict PCE, incorporating features such as energy levels and bandgap, while also offering insights into feature importance to guide material optimization. A significant case study curated a dataset of 111,000 molecules and trained an RF model using 2D and 3D molecular descriptors to predict LUMO and HOMO energy levels with an error below 0.16 eV, eliminating the need for DFT calculations and accelerating high-throughput screening of organic semiconductors for solar cell applications. Similarly, a cascaded SVM (CasSVM) model, a two-stage framework leveraging descriptors to predict key parameters including short-circuit current, open-circuit voltage, and fill factor, achieved a MAE of 0.35% for PCE predictions, approximately 10% of the average PCE value (3.89%), with an R^2 of 0.96, proving highly effective for pre-synthesis candidate evaluation. Neural networks, including convolutional neural networks, have delivered high accuracy (e.g., 91.02%) in predicting PCE for donor materials; however, limitations arise when trained on datasets like the Harvard Clean Energy Project (HCEP), where molecular diversity may not fully reflect real-world variability. A team compiled over 51,000 non-fullerene acceptors from HCEP, employing Gaussian process regression (GPR) to predict HOMO/LUMO levels and applying the Scharber model for PCE estimation. GPR has proven particularly valuable for property predictions in data-scarce settings, providing uncertainty estimates to inform experimental design, as demonstrated in studies optimizing synthesis sequences for OSC materials.

In the realm of morphology analysis, graph-based methods have been employed to investigate descriptors influencing photon absorption, exciton generation, and charge transport, as showcased in related research. Additionally, studies on non-fullerene acceptors from the Y6 series have integrated ML workflows to expedite material discovery, enhance device performance, and screen acceptor combinations for experimental validation. For ternary OSCs based on fullerene derivatives, RF models have attained R^2 scores of 0.77 by analyzing molecular fingerprints and descriptors. A comparison of simple constitutional descriptors with 3D structural descriptors for high-throughput screening revealed that incorporating 3D descriptors improved classification accuracy between high- and low-performing materials by approximately 15%. Decision trees and RFs have identified key structural features impacting PCE, while linear re-

gression models have elucidated relationships between molecular modifications and efficiency. Unsupervised techniques, such as *k*-means clustering, have grouped materials with similar properties to pinpoint promising families, and dimensionality reduction methods like principal component analysis (PCA) have simplified high-dimensional datasets by focusing on dominant factors. Semi-supervised learning has effectively balanced labeled and unlabeled data in settings with limited experimental data, while reinforcement learning approaches, including Q-learning and deep Q-networks, have optimized material synthesis and fabrication processes to improve stability and efficiency. Hybrid and multiscale modeling strategies integrate atomistic simulations with device-level models to connect molecular vibrations to macroscopic performance, as explored in various studies. Collectively, these developments underscore ML’s capacity to manage complex datasets, enhance predictive accuracy, and address data quality gaps, though challenges remain, including small dataset sizes, computational demands of 3D descriptors, and the need for robust feature engineering. By combining ML with quantum chemical simulations and high-throughput screening, these workflows have been refined, laying the groundwork for more stable and efficient OSCs suitable for commercial deployment.

The application of AI/ML in OSCs has evolved from initial proof-of-concept regressors on small experimental datasets to sophisticated, closed-loop frameworks targeting materials, processing, and device performance. Several public resources have significantly shaped this progress. Notably, the Harvard Organic Photovoltaics 2015 (HOSC15) dataset compiles experimental device records alongside corresponding quantum-chemical calculations, supporting supervised learning for HOMO/LUMO, energy gaps, and PCE proxies. Concurrently, the Clean Energy Project database (CEPDB) offers millions of in-silico candidates for virtual screening and pretraining, demonstrating the viability of high-throughput ML surrogates for organic semiconductors. Together, HOSC15 (experimental) and CEPDB (computational) serve as foundational resources anchoring much of the OSC ML literature.

In summary, early investigations relied on fixed descriptors, such as fingerprints and physicochemical features, along with classical ML techniques to predict PCE from donor structures gathered from the literature; later studies compared neural network models with tree-based and linear baselines, emphasizing the critical role of data curation and train/test split strategies in generalization performance. Recent reviews synthesize these trends, advocating for richer representations, uncertainty quantification, and standardized evaluation protocols, reflecting a maturing field that encompasses energy levels, absorption characteristics, and device performance metrics.

With the rise of graph learning, message-passing and attention-based models have become

the predominant methods for predicting molecular properties in OSCs. Beyond conventional atom-level graph representations, recent studies have enriched these models by integrating learned subunit and ring semantics, as well as multi-scale features, to more effectively capture essential chemical properties like conjugation and donor/acceptor electronic characteristics. In data-limited scenarios, multi-modal encoders that fuse learned graph features with language-derived descriptors from the OSC literature have yielded notable improvements in PCE prediction, underscoring the benefit of incorporating textual domain knowledge as a supplementary data source to enhance predictive capabilities.

The performance of OSC devices is significantly influenced by mesoscale morphology and processing conditions. Grazing-incidence wide-angle X-ray scattering (GIWAXS) and grazing-incidence X-ray scattering (GIXS) techniques have been widely used to derive quantitative descriptors of crystallinity and molecular orientation, and these features are increasingly integrated into ML workflows to bridge the gap between molecular structures, morphology, and device performance. For non-fullerene acceptors, recent research has uncovered strong correlations between core design, molecular packing, and processing history with both efficiency and stability, driving the development of ML models that explicitly account for morphology and processing metadata rather than depending solely on chemical structure. Recognizing that the power conversion efficiency of OSCs arises from donor-acceptor interactions and interfacial energetics, recent efforts have transitioned from single-molecule encoders to pair-aware models that concurrently learn representations of donor and acceptor components. These advanced models often combine graph-based features with physically grounded descriptors, such as energy level offsets and computed frontier orbital energies, to more precisely model the interfacial processes vital to OSC performance. Moreover, large-scale pretraining, paired with deep-learning-guided exploration of D-A pairs, has demonstrated the power of sophisticated representation learning to accelerate the discovery of promising candidates across vast chemical spaces.

The contributions of this thesis. This section integrates the findings from our four synergistic studies, robust *data cleansing* to ensure reliable labels, an automated ring/subunit representation (*Ring2Vec*), a ring-enhanced hierarchical transformer for single-molecule prediction (*RingFormer*), and a hierarchical graph transformer incorporating multi-level donor-acceptor interactions (*HGT-MI*), to develop a comprehensive end-to-end pipeline for precise molecular and device-level property prediction in organic solar cells.

From noisy labels to reliable supervision. A key challenge in training machine learning models for molecular property prediction in OSCs is the inherent noise in experimental labels, particularly for energy levels of non-fullerene acceptors. In the first work, we address this by

developing a robust data cleansing methodology that transforms unreliable cyclic voltammetry measurements into reliable supervision signals. The method leverages cross-calibration with more accurate device parameters, such as open-circuit voltage and optical bandgap, to estimate HOMO and LUMO levels indirectly. By defining discrepancies (ΔHOMO and ΔLUMO) between CV-derived and Voc-derived values, we systematically filter out noisy data points exceeding predefined thresholds, thereby enhancing dataset quality and reducing input errors that plague traditional density functional theory approaches, which often exhibit deviations of 0.2-0.5 eV. Complementing this cleansing step, we introduce a sub-unit-based molecular description, decomposing NFAs into aromatic sub-units and encoding their electronic properties as numerical inputs for machine learning models. This simplification not only mitigates computational complexity but also captures essential chemical features, enabling models like gradient boosting regression trees to achieve superior performance. The results demonstrate a remarkable reduction in prediction error to approximately 0.06 eV for HOMO/LUMO levels, far outperforming DFT while completing computations in mere seconds on standard hardware. This accuracy is validated through comparisons with experimental data, highlighting the method’s efficacy in low-data regimes typical of OSC research. The contributions of this work extend beyond methodological innovation, providing a catalyst for accelerated NFA development by enabling precise energy level control critical for optimizing donor-acceptor interactions and pushing OSC efficiencies toward 20%. By bridging noisy experimental labels with reliable supervision, the approach sets a precedent for data-driven strategies in materials science, potentially inspiring similar cleansing and representation techniques in other domains, such as drug discovery or polymer design, where experimental variability hinders predictive modeling.

Automated ring/subunit semantics with Ring2Vec. A significant limitation in molecular property prediction for organic solar cells is the manual labor required to encode high-level features like aromatic rings or subunits in non-fullerene acceptors, which traditional atom-level methods such as fingerprints or graph neural networks often overlook. In the second work, we introduce Ring2Vec, an innovative automated description method inspired by Word2Vec in natural language processing, treating aromatic rings as "words" and entire molecules as "sentences" to capture subunit semantics without expert labeling. The approach processes SMILES strings from the Harvard Clean Energy Project database to extract ring-based features, such as ring size, bond counts, heteroatoms, and fusions, at varying radii (e.g., single rings or fused pairs), pretraining a model to generate embeddings that encode contextual ring information. These are fused with atom-level embeddings from Fin2Vec (fingerprint-based) and MG2Vec (GNN-based) via self-attention, creating a multi-scale representation fed into machine learning models like

random forests for prediction. The results showcase Ring2Vec’s efficacy in predicting NFA energy levels, achieving a minimal mean absolute error of 0.06 eV for HOMO/LUMO, surpassing prior methods reliant on manual subunit labeling or DFT calculations, which typically incur errors of 0.2-0.5 eV and high computational costs. Validated on an experimental dataset, the framework enables rapid predictions (seconds on standard hardware), highlighting its superiority in capturing conjugation and electronic properties through automated, context-aware embeddings. This contribution advances OSC research by streamlining NFA design, facilitating precise energy level tuning for enhanced device efficiency, and reducing trial-and-error synthesis. More broadly, Ring2Vec’s label-free automation sets a versatile precedent for molecular representation in materials science, with potential extensions to drug discovery, polymer engineering, or other domains where subunit semantics drive property prediction, fostering efficient AI-driven innovation across chemical spaces.

Single-molecule hierarchical learning with RingFormer. A persistent challenge in predicting properties of organic solar cell molecules, such as power conversion efficiency, is the inadequate capture of hierarchical structural patterns, particularly intricate ring systems that dominate electron donor/acceptor behaviors, by conventional graph neural networks which excel at local atomic features but falter on global dependencies. In our third work, we introduce RingFormer, a specialized graph transformer that automates hierarchical learning for single-molecule OSC prediction by constructing a multi-level graph: an atom-level graph for atomic bonds, a ring-level graph for fused/non-fused ring interconnections, and an inter-level graph for cross-level relationships. The method employs message-passing GNNs for localized atom encoding, a novel cross-attention mechanism for global ring system patterns, and a hierarchical fusion strategy to integrate representations across levels, enabling expressive graph embeddings that reflect both local functional groups and high-level conjugated structures without manual feature engineering. Empirical evaluations on five curated OSC datasets, including the large-scale CEPDB, demonstrate RingFormer’s superior performance, outperforming 11 baselines with a 22.77% relative improvement over the closest competitor on CEPDB, underscoring its ability to model complex ring dependencies for accurate predictions of energy levels and PCE proxies. This contribution pioneers ring-enhanced graph transformers tailored for OSCs, bridging gaps in representation learning by automating hierarchical semantics, and holds promise for accelerating NFA/donor molecule screening to achieve efficiencies beyond 20%, with extensible applications to other conjugated systems in materials science like organic electronics or photovoltaics.

From molecules to pairs: multi-level D-A interactions with HGT-MI. Another critical bottleneck in organic solar cell development is the inability of traditional computational models to

capture the complex donor-acceptor interactions that dictate power conversion efficiency, a limitation rooted in their focus on isolated molecular properties rather than device-level dynamics. In our fourth study, we introduce the Hierarchical Graph Transformer with Multi-level Interactions, a novel framework that addresses this by modeling D-A interactions across atom, motif, and molecular levels. The method constructs a hierarchical graph representation, integrating atom-level bonds, motif-level structural units (e.g., aromatic rings, alkyl chains), and molecular-level D-A pairings, leveraging a multi-head attention mechanism to encode electronic couplings and energy-transfer dynamics. Local message-passing captures intra-molecular features, while global attention fuses inter-molecular interactions, enabling a comprehensive representation of interfacial processes critical to OSC performance. Experimental results on diverse OSC datasets demonstrate HGT-MI’s superior predictive power, achieving a significant reduction in mean absolute error for PCE predictions compared to state-of-the-art methods like DFT and single-molecule GNNs, with improvements exceeding 15% on benchmark datasets. This accuracy stems from its ability to model multi-scale interactions, validated through comparisons with experimental PCE values from device characterizations. The contribution of HGT-MI lies in its paradigm shift from single-molecule to pair-aware modeling, offering a scalable solution to accelerate OSC material discovery by optimizing D-A pairings without exhaustive synthesis. Its broad applicability extends to related systems like perovskite solar cells, light-emitting diodes, and batteries, where multi-level interactions govern performance, positioning HGT-MI as a transformative tool for computational materials science and paving the way for next-generation energy technologies.

Integrated interpretation. Drawing from the collective insights of these four studies, we have develop a seamless and comprehensive progression in molecular property prediction for organic solar cells, culminating in a robust, end-to-end pipeline that systematically addresses key challenges at every stage of the research process. The first component of this thesis addresses the inherent noise in experimental labels, particularly for non-fullerene acceptor energy levels, by developing a robust methodology that transforms unreliable cyclic voltammetry measurements into reliable supervision signals through cross-calibration with device parameters like open-circuit voltage and optical bandgap. This approach, which filters out discrepancies exceeding predefined thresholds, reduces prediction errors to an impressive 0.06 eV, far surpassing the 0.2-0.5 eV deviations typical of traditional density functional theory methods, and establishes a high-quality dataset foundation for subsequent modeling. Building on this, the Ring2Vec method introduces an automated, label-free approach inspired by natural language processing, extracting ring-based semantics from SMILES strings to enhance prediction accuracy and computational

efficiency, this was further refined through with the RingFormer framework. RingFormer employs a hierarchical graph transformer to integrate atom-level and ring-level features, leveraging message-passing and cross-attention mechanisms to achieve a 22.77% relative improvement in single-molecule predictions over existing baselines across five curated OSC datasets, including the large-scale CEPDB. Finally, the HGT-MI framework marks a paradigm shift by modeling multi-level donor-acceptor interactions across atom, motif, and molecular scales, using multi-head attention to encode electronic couplings and energy-transfer dynamics, resulting in over 15% better accuracy in PCE predictions compared to state-of-the-art methods like DFT and single-molecule GNNs. Together, these innovations from refining data quality through systematic cleansing, automating structural representations with contextual embeddings, enhancing hierarchical learning with expressive graph embeddings, to enabling pair-aware modeling that captures device-level interfacial processes, create a cohesive pipeline that accelerates OSC efficiency optimization, with the potential to push efficiencies beyond 20%. This integrated approach not only streamlines NFA and donor molecule screening but also offers a versatile and scalable blueprint for broader materials science applications, including drug discovery, polymer engineering, battery design, and beyond.

Discussion on PCE prediction. PCE represents a device-level performance metric that is fundamentally different from intrinsic molecular properties such as HOMO or LUMO energy levels. While the latter can often be reasonably approximated by single-molecule descriptors, PCE is jointly determined by the complex interplay between donor and acceptor materials, including their electronic compatibility, intermolecular interactions, and collective charge-transport behavior within the active layer. Consequently, accurate PCE prediction requires modeling strategies that move beyond isolated molecular representations and explicitly account for cross-molecular interactions. From a methodological perspective, the proposed interaction-aware framework is designed to address this challenge by jointly modeling donor and acceptor molecules and enabling information exchange across multiple structural levels. By incorporating atom-, motif-, and molecular-level interactions, the model captures both local chemical environments and higher-level structural compatibilities that are critical for device performance. This hierarchical design allows the model to learn interaction patterns that cannot be adequately represented by conventional single-molecule or purely atom-level approaches, thereby providing a more physically meaningful representation for PCE prediction.

Limitations. Despite the promising advances achieved through data cleansing, Ring2Vec, RingFormer, and HGT-MI, several limitations remain that warrant critical reflection. First, while the data cleansing framework substantially improves label reliability by cross-calibrating

CV with device-level parameters, it is still constrained by the availability and consistency of experimental measurements. The method cannot fully eliminate biases arising from diverse fabrication conditions, measurement protocols, or reporting standards across research groups. As such, systematic errors may persist, and generalization to datasets beyond OSCs may require domain-specific calibration strategies. Second, Ring2Vec and RingFormer, though effective in capturing hierarchical molecular semantics, are trained on datasets of limited diversity and size relative to the vast chemical design space. The scarcity of high-quality experimental data for non-fullerene acceptors and donor-acceptor pairs introduces risks of overfitting and may limit the models' robustness when extrapolating to novel chemistries, particularly those with unconventional backbones or side-chain engineering strategies not well represented in existing corpora. Third, while HGT-MI successfully models donor-acceptor interactions at multiple levels, it may not yet incorporate dynamic processes such as morphological evolution, phase separation, or exciton diffusion in device environments. These extrinsic factors critically influence OSC efficiency but remain outside the scope of purely molecular- and pair-level representations. Extending the framework to integrate morphological simulations or experimental imaging data remains an open challenge. Fourth, computational scalability and interpretability remain partial bottlenecks. Although the proposed models significantly outperform DFT in efficiency, large-scale pretraining of hierarchical graph transformers still demands considerable computational resources. Moreover, the embeddings produced by Ring2Vec, RingFormer, and HGT-MI, while highly predictive, remain partially opaque in their chemical interpretability, raising questions about their ability to provide actionable insights to experimental chemists beyond numerical accuracy. Finally, the pipeline is tailored specifically to organic solar cells, and while many of the principles, such as noise-robust data curation, subunit embeddings, and multi-level interaction modeling, are generalizable, applying them to other domains like drug discovery, polymers, or perovskites will likely require task-specific modifications, retraining, and validation.

Practical implications. Taken together, despite these limitations, the contributions of this thesis hold broad practical significance for advancing organic solar cell research and materials discovery more generally. By introducing a robust data cleansing framework, it provides the community with high-quality, reliable datasets that alleviate long-standing issues of noisy experimental measurements, thereby improving reproducibility and comparability across studies. The development of automated molecular representations through Ring2Vec and hierarchical modeling with RingFormer reduces dependence on manual feature engineering and domain expertise, allowing experimental chemists to rapidly evaluate structural design choices without requiring deep computational training. At the device level, the HGT-MI framework enables ac-

curate, pair-aware predictions of donor-acceptor performance, offering researchers and industry partners a powerful tool for virtual screening to guide synthesis toward high-efficiency materials while significantly cutting costs and time. More broadly, the efficiency gains achieved by these AI-driven models, outperforming traditional quantum-chemical calculations at a fraction of the computational expense, demonstrate a paradigm shift in computational materials science, offering a practical blueprint that can be adapted to related domains such as perovskites, OLEDs, and energy storage systems, where multi-level interactions and structural semantics are equally critical to performance.

Future Work. A major thrust of future research is to integrate rigorous three-dimensional (3D) molecular modeling into the pipeline, from single-molecule conformations to structure-aware design and generation of OSCs. Modern non-fullerene acceptors and high-performance donors feature large fused aromatic cores, conjugation-critical linkers, and long, flexible side chains that yield broad conformational ensembles and packing propensities. Preliminary tests with general-purpose 3D conversion tools (primarily tuned for drug-like chemotypes) show low success rates on OSC structures: they often fail to preserve aromatic planarity and conjugation continuity, mis-handle fused-ring rigidity, and collapse under side-chain sterics during relaxation. These observations motivate an OSC-specific 3D conversion model and, building on it, a 3D-aware generative framework that proposes synthesizable candidates together with plausible conformers optimized for electronic and morphological targets.

The first objective is an OSC-specialized 3D converter that maps annotated 2D graphs into high-fidelity *conformer ensembles* with calibrated uncertainty. Unlike generic distance-geometry routines, the converter will embed OSC-specific chemical and physical priors. Conjugation is treated as a first-class constraint: fused aromatic systems are modeled as *quasi-rigid* bodies with limited out-of-plane modes; linkers follow dihedral distributions that preserve conjugation; and side chains use realistic rotamer statistics that respect steric clashes near carbonyls, heteroatoms, and branching sites. Algorithmically, a hybrid approach is promising: (i) initialize with constraint-satisfying coordinates from an OSC-aware torsion/ring library (e.g., extending ETKDG with conjugation and aromaticity constraints, as well as E/Z, tautomer, and protonation-state handling); (ii) refine with an SE(3)-equivariant network (equivariant transformer or diffusion denoiser) that adjusts positions while honoring hard constraints on bond lengths/angles and ring closures; and (iii) apply lightweight energy guidance via learned surrogates of torsional and non-bonded terms to discourage unphysical distortions. The converter outputs *ensembles*, not single geometries, to capture multimodality in side-chain conformations and backbone twists that materially affect electronic coupling and packing.

Creating and sharing a high-quality OSC-3D reference set is essential for training and validation. We will curate: (a) experimentally determined crystal structures for representative donors/acceptors and oligomers; (b) quantum-optimized fragments spanning common linkers (e.g., thiophene–thiophene, benzothiadiazole-containing bridges, carbonyl/ester linkages) and side-chain attachment motifs; and (c) morphology-informed descriptors distilled from GIWAXS/SAXS (lamellar spacing, π – π distances, preferred backbone tilt). Because these measurements are process-dependent, descriptors will be conditioned on *fabrication variables* (e.g., solvent, annealing, substrate) to avoid confounding geometry with processing. To move beyond generic metrics such as RMSD, we will report OSC-aware evaluation criteria: aromatic plane deviation, conjugation continuity scores along the backbone, linker-specific dihedral histograms versus quantum references, clash counts in side-chain neighborhoods, and the stability of key surrogate properties (frontier-orbital estimates, reorganization energies, multipole moments) across the conformer ensemble. Charge-assignment protocols (e.g., RESP/AM1-BCC) will be standardized to ensure comparability of electrostatics-related features.

Uncertainty quantification will be built in from the start. Because many OSC properties depend on conformer averages, the converter will report calibrated dispersion (ensemble variance, evidential outputs, diffusion-score variance) and we will assess calibration with ECE/PICP/MIW metrics. Distribution-shift robustness will be tested using explicit out-of-distribution (OOD) challenge splits (novel fused cores/side-chain chemistries), with OOD detection based on embedding-space density/Mahalanobis scores combined with fingerprint distance thresholds.

With a reliable converter in place, the second objective is to make property prediction explicitly 3D-aware without relying on heavy *ab initio* pipelines. Fast surrogate features derived from conformer ensembles will capture aspects poorly represented in 2D: backbone planarity/curvature, conjugation length, side-chain steric indices, electrostatic multipoles, and coarse packing propensities (e.g., likelihood of face-on vs. edge-on in simple dimer configurations). Predictors will ingest *conformer-averaged features and their variances*, providing rapid estimates of energy-level alignment surrogates, reorganization energies, exciton-relevant couplings, and rudimentary miscibility indicators. Because solid-state levels differ from gas-phase values due to dielectric polarization and solid-state renormalization, environmental/process descriptors will be included explicitly when training against CV/device-level targets.

The third objective is a 3D-aware generative framework for OSC molecules. Rather than proposing candidates solely in 2D and hoping for realizable 3D structures, the generator will co-produce a chemically valid graph and one or more plausible 3D conformers consistent with

OSC priors. A hierarchical design space is natural: (1) propose fused-ring cores and fusion patterns from a motif library; (2) select linkers from a vetted catalog annotated with learned dihedral landscapes; (3) choose side-chain topologies (length, branching, heteroatom insertion, attachment positions) and, where appropriate, ionizable/zwitterionic substituents for stability or interfacial tuning; and (4) instantiate conformers through the OSC-specific converter with controlled stochasticity to explore packing-relevant microstates. Optimization will be multi-objective and risk-aware, targeting predicted frontier-level offsets, small reorganization energies, planarity/stacking proxies, blend-compatibility hints, photostability heuristics, and practical constraints such as synthetic accessibility (SA score/SCScore or retrosynthesis filters), toxicity flags, and cost proxies. To avoid reward hacking of surrogates, we will (i) penalize high conformer-to-conformer variance, (ii) enforce strict validity checks from the converter (conjugation continuity, clash thresholds, torsion bounds), (iii) apply adversarial stress tests (Randomized SMILES, conformer perturbations), and (iv) require that “winners” pass an independent physics kernel before promotion.

Integrating this generation loop into an experiment-in-the-loop workflow is a key translational step. We envision active-learning cycles that propose small batches with uncertainty- and diversity-aware acquisition (e.g., UCB with variance penalty, BALD, qEHVI), synthesize and characterize a subset, update torsion priors/converter/surrogates with the new data, and iterate. Even partial feedback-GIWAXS peak positions, CV gaps, simple solubility/processability read-outs, will refine priors on planarity, π - π spacing, and side-chain behavior, improving realism for both conversion and generation. Over time, the system accumulates OSC-specific knowledge that broadens coverage of unusual heteroacenes, non-intuitive linkers, and side-chain architectures.

For community impact and reproducibility, we will release a modular OSC-3D toolkit: (i) a callable conformer generator with OSC-aware constraints and equivariant refinement; (ii) feature extractors for ensemble-based 3D descriptors; (iii) fast property surrogates suitable for screening; (iv) a 3D-aware generator with multi-objective controllers (Bayesian optimization, evolutionary search, or RL in latent spaces); and (v) standardized benchmarks/leaderboards reporting not only prediction accuracy but also geometric validity, calibration, stability across conformers, and diversity/novelty under synthetic constraints. We will publish torsion-scan libraries, conformer sets, scripts, random seeds, and *challenge splits* that emphasize extrapolation to unfamiliar cores or side-chain chemistries.

Several technical questions will guide the agenda: How to balance hard chemistry rules with learned flexibility so the converter is both valid and expressive? What is the minimum quantum/experimental supervision needed to learn trustworthy torsion/planarity priors, and

how should multi-fidelity training (semi-empirical with selective high-level calculations) be scheduled for scalability? How do we design uncertainty estimators that remain calibrated under distribution shift as the generator proposes increasingly novel structures? Which proposal mechanisms most efficiently explore conformational and substituent spaces without exhaustive enumeration, especially for long side chains and heavily fused cores? And how can we quantify morphology relevance with lightweight descriptors that correlate with device performance while keeping the pipeline fast enough for thousands to millions of design iterations?

Finally, we anticipate risks and propose mitigations. Out-of-distribution structures (e.g., exotic heteroatom placements or extreme curvature) may break learned priors; the converter will detect OOD motifs via embedding density/fingerprint distance and trigger constrained quantum refinement only for the offending fragments, caching results to expand coverage. Surrogates may be biased by training data; we will employ ensembling, conformer-averaged targets, and periodic realignment to experimental readouts. Generators can exploit surrogate weaknesses; we will pair adversarial screening with hard chemistry checks, diversity regularization, and independent validation stacks before promotion.

In sum, future work will make 3D geometry a core asset rather than an afterthought: a robust OSC-specific converter that respects conjugation and fused-ring physics; fast 3D-aware surrogates that translate geometry into screening power; and a principled, uncertainty-aware generator that proposes synthesizable, high-promise candidates under realistic constraints. This trajectory directly addresses the low success rate of existing 3D tools on OSC molecules and opens a practical path from structural realism to accelerated discovery.

References

- [1] Elyas Abbasi Jannat Abadi, Harikrishna Sahu, Seyed Morteza Javadpour, and Masoud Goharimanesh. Interpretable machine learning for developing high-performance organic solar cells. *Materials Today Energy*, 25:100968, 2022.
- [2] Md Akhtaruzzaman and Vidhya Selvanathan. *Comprehensive guide on organic and inorganic solar cells: fundamental concepts to fabrication methods*. Academic Press, 2021.
- [3] Michael P. Allen and Dominic J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, 2nd edition, 2017.
- [4] Kaikai An, Wenkai Zhong, Feng Peng, Wanyuan Deng, Ying Shang, Huilei Quan, Hong Qiu, Cheng Wang, Feng Liu, Hongbin Wu, Ning Li, Fei Huang, and Lei Ying. Mastering morphology of non-fullerene acceptors towards long-term stable organic solar cells. *Nature Communications*, 14(1):2688, 2023.
- [5] Robin Basu, Fabian Gumpert, Jan Lohbreier, Pierre-Olivier Morin, Varun Vohra, Yang Liu, Yinhua Zhou, Christoph J. Brabec, Hans-Joachim Egelhaaf, and Andreas Distler. Large-area organic photovoltaic modules with 14.5% certified efficiency. *Joule*, 8:147792, 2024.
- [6] Vinayak Bhat, Parker Sornberger, Balaji Sesha Sarath Pokuri, Rebekah Duke, Baskar Ganapathysubramanian, and Chad Risko. Electronic, redox, and optical property prediction of organic π -conjugated molecules through a hierarchy of machine learning approaches. *Chemical Science*, 14:203–213, 2023.
- [7] Abhiroop Bhattacharya and Sylvain G. Cloutier. Matmmfuse: Multi-modal fusion model for material property prediction. *arXiv preprint arXiv:2505.04634*, 2025.
- [8] Pengqing Bi, Shaoqing Zhang, Junzhen Ren, Zhihao Chen, Zhong Zheng, Yong Cui, Jianqiu Wang, Shijie Wang, Tao Zhang, Jiayao Li, Ye Xu, Jinzhao Qin, Cunbin An, Wei Ma, Xiaotao Hao, and Jianhui Hou. A high-performance nonfused wide-bandgap acceptor for versatile photovoltaic applications. *Advanced Materials*, 34(5), 2021.
- [9] Xavier Bresson and Thomas Laurent. Residual gated graph convnets. *arXiv preprint arXiv:1711.07553*, 2017.

-
- [10] Kieron Burke. Perspective on density functional theory. *The Journal of Chemical Physics*, 136(15):150901, 2012.
- [11] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.
- [12] Guilong Cai, Yuhao Li, Jiayu Wang, Yiqun Zhang, Xinhui Lu, Jiarong Lian, Pengju Zeng, Yiping Wang, and Xiaowei Zhan. Effects of π -bridge on fused-ring electron acceptor dimers. *ACS Applied Polymer Materials*, 3(1):23–29, 2020.
- [13] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.
- [14] Julio C. Carrillo-Sendejas and José-Luis Maldonado. Progress in organic solar cells: Materials, challenges, and novel strategies for niche applications. *APL Energy*, 3(2):021501, 2025.
- [15] Anand Chandrasekaran, Deepak Kamal, Rohit Batra, Chiho Kim, Lihua Chen, and Rampi Ramprasad. Solving the electronic structure problem with machine learning. *npj Computational Materials*, 2019.
- [16] Chen Chen, Liang Wang, Weiyi Xia, Ke Qiu, Chuanhang Guo, Zirui Gan, Jing Zhou, Yuandong Sun, Dan Liu, Wei Li, and Tao Wang. Molecular interaction induced dual fibrils towards organic solar cells with certified efficiency over 20%. *Nature Communications*, 15:6865, 2024.
- [17] Xian-Kai Chen, Veaceslav Coropceanu, and Jean-Luc Brédas. Assessing the nature of the charge-transfer electronic states in organic solar cells. *Nature Communications*, 9:5295, 2018.
- [18] Yuzhong Chen, Fujin Bai, Zhengxing Peng, Lei Zhu, Jianquan Zhang, Xinhui Zou, Yunpeng Qin, Ha Kyung Kim, Jun Yuan, Lik-Kuen Ma, Jie Zhang, Han Yu, Philip C. Y. Chow, Fei Huang, Yingping Zou, Harald Ade, Feng Liu, and He Yan. Asymmetric alkoxy and alkyl substitution on nonfullerene acceptors enabling high-performance organic solar cells. *Advanced Energy Materials*, 11(3):2003141, 2021.
- [19] Zhengdao Chen, Lei Chen, Soledad Villar, and Joan Bruna. Can graph neural networks count substructures? *Advances in neural information processing systems*, 33:10383–10395, 2020.

- [20] Fuliang Cheng, Yongjie Cui, Feng Ding, Zeng Chen, Qian Xie, Xinxin Xia, Peipei Zhu, Xinhui Lu, Haiming Zhu, Xunfan Liao, and Yiwang Chen. Terpolymerization and regioisomerization strategy to construct efficient terpolymer donors enabling high-performance organic solar cells. *Advanced Materials*, 35(30), 2023.
- [21] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.
- [22] Martin Cigánek, Jan Richtár, Martin Weiter, and Jozef Krajčovič. Organic π -conjugated molecules: From nature to artificial applications—where are the boundaries? *Israel Journal of Chemistry*, 62, 2021.
- [23] Connor W. Coley, William H. Green, and Klavs F. Jensen. Machine learning in computer-aided synthesis planning. *Accounts of Chemical Research*, 51(5):1281–1289, 2018.
- [24] Felix Creutzig, Peter Agoston, Jan Christoph Goldschmidt, Gunnar Luderer, Gregory Nemet, and Robert C. Pietzcker. The underestimated potential of solar energy to mitigate climate change. *Nature Energy*, 2, 2017.
- [25] Yong Cui, Ye Xu, Huifeng Yao, Pengqing Bi, Ling Hong, Jianqi Zhang, Yunfei Zu, Tao Zhang, Jinzhao Qin, Junzhen Ren, Zhihao Chen, Chang He, Xiaotao Hao, Zhixiang Wei, and Jianhui Hou. Single-junction organic photovoltaic cell with 19% efficiency. *Advanced Materials*, 33(41), 2021.
- [26] Yongjie Cui, Peipei Zhu, Xunfan Liao, and Yiwang Chen. Recent advances of computational chemistry in organic solar cell research. *Journal of Materials Chemistry C*, 8(45):15920–15939, 2020.
- [27] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3:1503–1507, 2008.
- [28] Jianyuan Deng, Zhibo Yang, Hehe Wang, Iwao Ojima, Dimitris Samaras, and Fusheng Wang. A systematic study of key elements underlying molecular property prediction. *Nature Communications*, 14:6395, 2023.
- [29] Min Deng, Huifeng Meng, Xiaopeng Xu, Jie Tang, Liyang Yu, Ruipeng Li, and Qiang Peng. Unique w-shape y6 isomer as effective solid additive for high-performance pm6:y6 polymer solar cells. *Chemical Engineering Journal*, 440, 2022.

- [30] Pengfei Ding, Daobin Yang, Shuncheng Yang, and Ziyi Ge. Stability of organic solar cells: toward commercial applications. *Chemical Society Reviews*, 53, 2024.
- [31] Zhihao Ding, Ting Zhang, Yiran Li, Jieming Shi, and Chen Jason Zhang. Ringformer: A ring-enhanced graph transformer for organic solar cell property prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 155–163. AAAI Press, 2025.
- [32] Joseph L Durant, Burton A , Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002.
- [33] David K. Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, pages 2224–2232, 2015.
- [34] Andreas Eibeck, Daniel Nurkowski, Angiras Menon, Jiaru Bai, Jinkui Wu, Li Zhou, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. Predicting power conversion efficiency of organic photovoltaics: Models and data analysis. *ACS Omega*, 6(37):23764–23775, 2021.
- [35] Benedek Fabian, Thomas Edlich, Hélène Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.
- [36] Fraunhofer ISE. Photovoltaics report. Technical report, Fraunhofer Institute for Solar Energy Systems, 2025.
- [37] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 2nd edition, 2001.
- [38] Hongyang Gao and Shuiwang Ji. Graph u-nets. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2083–2092. PMLR, 2019.
- [39] Huanhuan Gao, Chenyang Han, Xiangjian Wan, and Yongsheng Chen. Recent progress in non-fused ring electron acceptors for high-performance organic solar cells. *Industrial Chemistry & Materials*, 1(1):60–78, 2023.

- [40] Jian Gao, Yaokai Li, Shuixing Li, Xinxin Xia, Xinhui Lu, Minmin Shi, and Hongzheng Chen. Non-fullerene acceptors with nitrogen-containing six-membered heterocycle cores for the applications in organic solar cells. *Solar Energy Materials and Solar Cells*, 225, 2021.
- [41] Wei Gao, Xiaoling Ma, Qiaoshi An, Jinhua Gao, Cheng Zhong, Fujun Zhang, and Chuluo Yang. An asymmetrical fused-ring electron acceptor designed by a cross-conceptual strategy achieving 15.6% efficiency. *Journal of Materials Chemistry A*, 8(29):14583–14591, 2020.
- [42] Wei Gao, Feng Qi, Zhengxing Peng, Francis R. Lin, Kui Jiang, Cheng Zhong, Werner Kaminsky, Zhiqiang Guan, Chun-Sing Lee, Tobin J. Marks, Harald Ade, and Alex K.-Y. Jen. Achieving 19% power conversion efficiency in planar-mixed heterojunction organic solar cells using a pseudosymmetric electron acceptor. *Advanced Materials*, 34(32), 2022.
- [43] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1263–1272, 2017.
- [44] Stefan Goedecker. Linear scaling electronic structure methods. *Reviews of Modern Physics*, 71:1085–1123, 1999.
- [45] Thavamani Gokulnath, Jungmin Choi, Ho-Yeol Park, Kyungmin Sung, Yeongju Do, Hyungjin Park, Junyoung Kim, Saripally Sudhaker Reddy, Jehan Kim, Myungkwan Song, Jinhwan Yoon, and Sung-Ho Jin. A wide-bandgap π -conjugated polymer for high-performance ternary organic solar cells with an efficiency of 17.40%. *Nano Energy*, 89:106407, 2021.
- [46] Fuchun Gou, Zhu Ma, Qiang Yang, Hao Du, Yixian Li, Qian Zhang, Wei You, Yi Chen, Zhuowei Du, Junbo Yang, Nan He, Junxin Luo, Zichen Liu, Zilu Tian, Maozhu Mao, Kai Liu, Jian Yu, Anan Zhang, Fan Min, Kuan Sun, and Ningqiang Xuan. Machine learning-assisted prediction and control of bandgap for organic-inorganic metal halide perovskites. *ACS Applied Materials & Interfaces*, 17:18383–18393, 2025.
- [47] JS Griffith and LE Orgel. Ligand-field theory. *Quarterly Reviews, Chemical Society*, 11(4):381–393, 1957.

- [48] Zhanglin Guo, Ajay Kumar Jena, Gyu Min Kim, and Tsutomu Miyasaka. The high open-circuit voltage of perovskite solar cells: A review. *Energy & Environmental Science*, 15(8):3171–3222, 2022.
- [49] Zhichun Guo, Kehan Guo, Bozhao Nan, Yijun Tian, Roshni G. Iyer, Yihong Ma, Olaf Wiest, Xiangliang Zhang, Wei Wang, Chuxu Zhang, and Nitesh V. Chawla. Graph-based molecular representation learning. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6638–6646. International Joint Conferences on Artificial Intelligence Organization, 8 2023. Survey Track.
- [50] Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S Sánchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M Brockway, and Alán Aspuru-Guzik. The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *The Journal of Physical Chemistry Letters*, 2(17):2241–2251, 2011.
- [51] Johannes Hachmann, Roberto Olivares-Amaya, Adrian Jinich, Anthony L. Appleton, Martin A. Blood-Forsythe, László R. Seress, Carolina Roman-Salgado, Kai Trepte, Sule Atahan-Evrenk, Suleyman Er, Supriya Shrestha, Rajib Mondal, Anatoliy Sokolov, Zhenan Bao, and Alan Aspuru-Guzik. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry-the harvard clean energy project. *Energy & Environmental Science*, 7:698–704, 2014.
- [52] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1025–1035, 2017.
- [53] Xiaoxin He, Bryan Hooi, Thomas Laurent, Adam Perold, Yann LeCun, and Xavier Bresson. A generalization of vit/mlp-mixer to graphs. In *Proceedings of the 40th International Conference on Machine Learning*, pages 12724–12745. PMLR, 2023.
- [54] Rebecca R. Hernandez, Alona Armstrong, Jennifer Burney, Greer Ryan, Kara Moore-O’Leary, Ibrahima Diédhiou, Steven M. Grodsky, Leslie Saul-Gershenz, Rob Davis, Jordan Macknick, Dustin Mulvaney, Garvin A. Heath, Shane B. Easter, Madison K. Hoffacker, Michael F. Allen, and Daniel M. Kammen. Techno-ecological synergies of solar energy for global sustainability. *Nature Sustainability*, 2:560–568, 2019.

- [55] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136:B864–B871, 1964.
- [56] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical Review*, 136:B864–B871, 1964.
- [57] Shion Honda, Shoi Shi, and Hiroki R. Ueda. Smiles transformer: Pre-trained molecular fingerprint for low-data drug discovery. *arXiv preprint arXiv:1911.04738*, 2019.
- [58] Jianhui Hou, Olle Inganäs, Richard H Friend, and Feng Gao. Organic solar cells based on non-fullerene acceptors. *Nature materials*, 17(2):119–128, 2018.
- [59] Ran Hou, Miao Li, Xueqing Ma, Hao Huang, Hao Lu, Qingqing Jia, Yahui Liu, Xinjun Xu, Hai-Bei Li, and Zhishan Bo. Noncovalently fused-ring electron acceptors with c2v symmetry for regulating the morphology of organic solar cells. *ACS Applied Materials & Interfaces*, 12(41):46220–46230, 2020.
- [60] Weihua Hu, Bowen Liu, José Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *ICLR*, 2020.
- [61] Shaorong Huang, Yufei Wang, Lintao Hou, and Lie Chen. Large-area flexible organic solar cells. *npj Flexible Electronics*, 5:30, 2021.
- [62] Chaudhery Mustansar Hussain. *Handbook of nanomaterials for industrial applications*. Elsevier, 2018.
- [63] Riaz Hussain, Muhammad Adnan, Kainat Atiq, Muhammad Usman Khan, Zahoor H. Farooqi, Javed Iqbal, and Robina Begum. Designing of silolothiophene-linked triphenylamine-based hole transporting materials for perovskites and donors for organic solar cells-a dft study. *Solar Energy*, 253:187–198, 2023.
- [64] Raymond P. Iczkowski and John L. Margrave. Electronegativity. *Journal of the American Chemical Society*, 83(17):3547–3551, 1961.
- [65] International Energy Agency. Solar pv. Technical report, International Energy Agency (IEA), 2025. Accessed 2025-09-01.
- [66] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: A pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3:015022, 2022.

- [67] Clemens Isert, Kenneth Atz, and Gisbert Schneider. Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology*, 79:102548, 2023.
- [68] Arnulf Jäger-Waldau. Snapshot of photovoltaics - february 2024. *EPJ Photovoltaics*, 15:1–24, 2024.
- [69] Muhammad Ramzan Saeed Ashraf Janjua. Photovoltaic properties and enhancement in near-infrared light absorption capabilities of acceptor materials for organic solar cell applications: A quantum chemical perspective via dft. *Journal of Physics and Chemistry of Solids*, 171:110970, 2022.
- [70] Muhammad Ramzan Saeed Ashraf Janjua, Ahmad Irfan, Mohamed Hussien, Muhammad Ali, Muhammad Saqib, and Muhammad Sulaman. Machine-learning analysis of small-molecule donors for fullerene based organic solar cells. *Energy Technology*, 10:2200019, 2022.
- [71] Ke Jin, Zuo Xiao, and Liming Ding. 18.69% pce from organic solar cells. *Journal of Semiconductors*, 42(6):060502, 2021.
- [72] Stuart Warren Jonathan Clayden, Nick Greeves. *Organic chemistry*. Oxford University Press, 2012.
- [73] Wei Ju, Zequn Liu, Yifang Qin, Bin Feng, Chen Wang, Zhihui Guo, Xiao Luo, and Ming Zhang. Few-shot molecular property prediction via hierarchically structured learning on relation graphs. *Neural Networks*, 163:122–131, 2023.
- [74] Jiye Kim, Seungbeom Lee, Dongwoo Kim, Sungsoo Ahn, and Jaesik Park. Substructure-atom cross attention for molecular representation learning. In *Advances in Neural Information Processing Systems*, 2022.
- [75] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2016.
- [76] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140:A1133–A1138, 1965.
- [77] Oliver Kramer. Scikit-learn. *Machine learning for evolution strategies*, pages 45–53, 2016.
- [78] Kakaraparthi Kranthiraja and Akinori Saeki. Experiment-oriented machine learning of polymer:non-fullerene organic solar cells. *Advanced Functional Materials*, 31, 2021.

- [79] Mario Krenn, Florian Häse, Akshat Kumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.
- [80] Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems*, 34:21618–21629, 2021.
- [81] Lukas Kruitwagen, K. T. Story, Johannes Friedrich, Laura Byers, Sarah Skillman, and Cameron Hepburn. A global inventory of photovoltaic solar energy generating units. *Nature*, 598:604–610, 2021.
- [82] Hanjian Lai, Meigen Guo, Yulin Zhu, Lin Chen, Pu Tan, Chuluo Yang, and Feng He. The cis- and trans-orientation of benzo[1,2-b:4,5-b']dithiophene-based isomers in organic solar cells. *Materials Chemistry Frontiers*, 5(3):1486–1494, 2021.
- [83] Greg Landrum. Rdkit: Open-source cheminformatics software. <https://www.rdkit.org/>, 2016.
- [84] Jaewan Lee, Changyoung Park, Hongjun Yang, Sungbin Lim, Woohyung Lim, and Sehui Han. Cast: Cross attention based multimodal fusion of structure and text for materials property prediction. *arXiv preprint arXiv:2502.06836*, 2025.
- [85] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3734–3743. PMLR, 2019.
- [86] Min-Hsuan Lee. Identifying correlation between the open-circuit voltage and the frontier orbital energies of non-fullerene organic solar cells based on interpretable machine-learning approaches. *Solar Energy*, 234:360–367, 2022.
- [87] Min-Hyung Lee. Robust random forest based non-fullerene organic solar cells efficiency prediction. *Organic Electronics*, 76:105504, 2020.
- [88] Lucia Leonat, Gabriela Sbarcea, and Ioan Viorel Branzoi. Cyclic voltammetry for energy levels estimation of organic materials. *U.P.B. Sci. Bull., Series B*, 75(3):111–118, 2013.
- [89] Chao Li, Jiali Song, Hanjian Lai, Huotian Zhang, Rongkun Zhou, Jinqiu Xu, Haodong Huang, Liming Liu, Jiabin Gao, Yuxuan Li, Min Hun Jee, Zilong Zheng, Sha Liu, Jun Yan, Xian-Kai Chen, Zheng Tang, Chen Zhang, Han Young Woo, Feng He, Feng

- Gao, He Yan, and Yanming Sun. Non-fullerene acceptors with high crystallinity and photoluminescence quantum yield enable >20% efficiency organic solar cells. *Nature Materials*, 24:433–443, 2025.
- [90] Chao Li, Jiadong Zhou, Jiali Song, Jinqiu Xu, Huotian Zhang, Xuning Zhang, Jing Guo, Lei Zhu, Donghui Wei, Guangchao Han, Jie Min, Yuan Zhang, Zengqi Xie, Yuanping Yi, He Yan, Feng Gao, Feng Liu, and Yanming Sun. Non-fullerene acceptors with branched side chains and improved molecular packing to exceed 18% efficiency in organic solar cells. *Nature Energy*, 6(6):605–613, 2021.
- [91] Jing Li, Jiage Song, Honggang Chen, Qingya Wei, Jun Yuan, Hongjian Peng, and Yingping Zou. Regulating the properties of small molecular acceptors with different end groups. *Solar Energy*, 223:100–105, 2021.
- [92] Shitong Li, Zhixiang Li, Xiangjian Wan, and Yongsheng Chen. Recent progress in flexible organic solar cells. *Materials Today Physics*, 33:100951, 2023.
- [93] Shuixing Li, Lingling Zhan, Wenchao Zhao, Shuhua Zhang, Basharat Ali, Zhisheng Fu, Tsz-Ki Lau, Xinhui Lu, Minmin Shi, Chang-Zhi Li, Jianhui Hou, and Hongzheng Chen. Revealing the effects of molecular packing on the performances of polymer solar cells based on a-d-c-d-a type non-fullerene acceptors. *Journal of Materials Chemistry A*, 6(25):12132–12141, 2018.
- [94] Sunsun Li, Long Ye, Wenchao Zhao, Shaoqing Zhang, Subhrangsu Mukherjee, Harald Ade, and Jianhui Hou. Energy-level modulation of small-molecule electron acceptors to achieve over 12% efficiency in polymer solar cells. *Advanced Materials*, 28(42):9423–9429, 2016.
- [95] Xian’e Li, Qilun Zhang, Jianwei Yu, Ye Xu, Rui Zhang, Chuanfei Wang, Huotian Zhang, Simone Fabiano, Xianjie Liu, Jianhui Hou, Feng Gao, and Mats Fahlman. Mapping the energy level alignment at donor/acceptor interfaces in non-fullerene organic solar cells. *Nature Communications*, 13(1):2046, 2022.
- [96] Zhixiang Li, Min Jiang, Shan Wang, and Shu Zhang. Deep learning methods for molecular representation and property prediction. *Drug Discovery Today*, 27:103373, 2022.
- [97] Youcai Liang, Difei Zhang, Zerun Wu, Tao Jia, Larry Lüer, Haoran Tang, Ling Hong, Jiabin Zhang, Kai Zhang, Christoph J. Brabec, Ning Li, and Fei Huang. Organic solar

- cells using oligomer acceptors for improved stability and efficiency. *Nature Energy*, 7:1180–1190, 2022.
- [98] Delong Liu, Liyan Yang, Yang Wu, Xiaohui Wang, Yan Zeng, Guangchao Han, Huifeng Yao, Sunsun Li, Shaoqing Zhang, Yun Zhang, Yuanping Yi, Chang He, Wei Ma, and Jianhui Hou. Tunable electron donating and accepting properties achieved by modulating the steric hindrance of side chains in a-d-a small-molecule photovoltaic materials. *Chemistry of Materials*, 30(3):619–628, 2018.
- [99] Feng Liu, Liang Zhou, Wenrui Liu, Zichun Zhou, Qihui Yue, Wenyu Zheng, Ri Sun, Wuyue Liu, Shengjie Xu, Haijun Fan, Liheng Feng, Yuanping Yi, Wenkai Zhang, and Xiaozhang Zhu. Organic solar cells with 18% efficiency enabled by an alloy acceptor: A two-in-one strategy. *Advanced Materials*, 33:2100830, 2021.
- [100] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nature Machine Intelligence*, 2023.
- [101] Wenxu Liu, Weiping Li, Jiannian Yao, and Chuanlang Zhan. Achieving high short-circuit current and fill-factor via increasing quinoidal character on nonfullerene small molecule acceptor. *Chinese Chemical Letters*, 29(3):381–384, 2018.
- [102] Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [103] Salvador A. Lopez, Benjamin Sanchez-Lengeling, Joao de Goes Soares, and Alán Aspuru-Guzik. Design principles and top non-fullerene acceptor candidates for organic photovoltaics. *Joule*, 1:857–870, 2017.
- [104] Steven A. Lopez, Edward O. Pyzer-Knapp, Gregor N. Simm, Trevor Lutzow, Kewei Li, Laszlo R. Seress, Johannes Hachmann, and Alán Aspuru-Guzik. The harvard organic photovoltaic dataset. *Scientific Data*, 3:160086, 2016.
- [105] Andreas Loukas. What graph neural networks cannot learn: depth vs width. In *International Conference on Learning Representations*, 2019.

- [106] Zhenghui Luo, Ruijie Ma, Tao Liu, Jianwei Yu, Yiqun Xiao, Rui Sun, Guanshui Xie, Jun Yuan, Yuzhong Chen, Kai Chen, Gaoda Chai, Huiliang Sun, Jie Min, Jian Zhang, Yingping Zou, Chuluo Yang, Xinhui Lu, Feng Gao, and He Yan. Fine-tuning energy levels via asymmetric end groups enables polymer solar cells with efficiencies over 17%. *Joule*, 4:1236–1247, 2020.
- [107] Asif Mahmood and Jin-Liang Wang. Machine learning for high performance organic solar cells: Current scenario and future prospects. *Energy & Environmental Science*, 14(1):90–105, 2021.
- [108] Prateek Malhotra, Subhayan Biswas, and Ganesh D. Sharma. Directed message passing neural network for predicting power conversion efficiency in organic solar cells. *ACS Applied Materials & Interfaces*, 15:37741–37747, 2023.
- [109] Aurelio Mateo-Alonso. π -conjugated materials: Here, there, and everywhere. *Chemistry of Materials*, 35:1467–1469, 2023.
- [110] Krzysztof Maziarz, Henry Jackson-Flux, Pashmina Cameron, Finton Sirockin, Nadine Schneider, Nikolaus Stiefl, Marwin Segler, and Marc Brockschmidt. Learning to extend molecular scaffolds with structural motifs. <https://arxiv.org/abs/2103.03864>, 2021.
- [111] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [112] Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gwoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624:80–86, 2023.
- [113] Murat Mesta, Jin Hyun Chang, Suranjan Shil, Kristian Sommer Thygesen, and Juan Maria García Lastra. A protocol for fast prediction of electronic and optical properties of donor-acceptor polymers using density functional theory and the tight-binding method. *The Journal of Physical Chemistry A*, 123(23):4980–4989, 2019.
- [114] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeff Dean. Efficient estimation of word representations in vector space. <https://arxiv.org/abs/1301.3781>, 2013.
- [115] Yuta Miyake and Akinori Saeki. Machine learning-assisted development of organic solar cell materials: Issues, analyses, and outlooks. *The Journal of Physical Chemistry Letters*, 12:12391–12401, 2021.

- [116] Kathleen Isabelle Moineau-Chane Ching. Impact of Alkyl-Based Side Chains in Conjugated Materials for Bulk Heterojunction Organic Photovoltaic Cells-A Review. *Energies*, 16(18):6639, 2023.
- [117] Gareth John Moore, Olivier Bardagot, and Natalie Banerji. Deep transfer learning: A fast and accurate tool to predict the energy levels of donor molecules for organic photovoltaics. *Advanced Theory and Simulations*, 5(5), 2022.
- [118] Viggo Moro, Charlotte Loh, Rumen Dangovski, Ali Ghorashi, Andrew Ma, Zhuo Chen, Samuel Kim, Peter Y. Lu, Thomas Christensen, and Marin Soljacic. Multimodal foundation models for material property prediction and discovery. *Newton*, 1:100016, 2025.
- [119] Shinji Nagasawa, Eman Al-Naamani, and Akinori Saeki. Computer-aided screening of conjugated polymers for organic solar cells: Classification by random forest. *The Journal of Physical Chemistry Letters*, 9(10):2639–2646, 2018.
- [120] NREL. Best research-cell efficiencies (data guide). Technical report, National Renewable Energy Laboratory (NREL), July 2025.
- [121] NREL. Best research-cell efficiency chart (interactive). Technical report, National Renewable Energy Laboratory (NREL), 2025. Accessed 2025-09-01.
- [122] Chandramouli Nyshadham, Matthias Rupp, Brayden Bekker, Alexander V. Shapeev, Tim Mueller, Conrad W. Rosenbrock, Gábor Csányi, David W. Wingate, and Gus L. W. Hart. Machine-learned multi-system surrogate models for materials prediction. *npj Computational Materials*, 2019.
- [123] Roberto Olivares-Amaya, Carlos Amador-Bedolla, Johannes Hachmann, Sule Atahan-Evrenk, Roel S. Sanchez-Carrera, Leslie Vogt, and Alan Aspuru-Guzik. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy & Environmental Science*, 4:4862–4869, 2011.
- [124] Felipe Oviedo, Zekun Ren, Shijing Sun, Charles Settens, Zhe Liu, Noor Titan Putri Hartono, Savitha Ramasamy, Brian L. DeCost, Siyu I. P. Tian, Giuseppe Romano, Aaron Gilad Kusne, and Tonio Buonassisi. Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials*, 5:60, 2019.

- [125] Shuting Pang, Xia Zhou, Song Zhang, Haoran Tang, Sujata Dhakal, Xiaodan Gu, Chunhui Duan, Fei Huang, and Yong Cao. Nonfused nonfullerene acceptors with an a-d-a'-d-a framework and a benzothiadiazole core for high-performance organic solar cells. *ACS Applied Materials & Interfaces*, 12(14):16531–16540, 2020.
- [126] Arindam Paul, Alona Furmanchuk, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Property prediction of organic donor molecules for photovoltaic applications using extremely randomized trees. *Molecular Informatics*, 38:1900038, 2019.
- [127] Arindam Paul, Dipendra Jha, Reda Al-Bahrani, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Transfer learning using ensemble neural networks for organic solar cell screening. <https://arxiv.org/abs/1903.03178>, 2019.
- [128] Souvik Pore, Arkaprava Banerjee, and Kunal Roy. Application of machine learning-based read-across structure-property relationship (raspr) as a new tool for predictive modelling: Prediction of power conversion efficiency (pce) for selected classes of organic dyes in dye-sensitized solar cells (dsscs). *Molecular Informatics*, 43:e202300210, 2024.
- [129] Edward O. Pyzer-Knapp, Jed W. Pitera, Peter W. J. Staar, Seiji Takeda, Teodoro Laino, Daniel P. Sanders, James Sexton, John R. Smith, and Alessandro Curioni. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, 8(1):84, 2022.
- [130] Deping Qian, Zilong Zheng, Huifeng Yao, Wolfgang Tress, Thomas R. Hopper, Shula Chen, Sunsun Li, Jing Liu, Shangshang Chen, Jiangbin Zhang, Xiao-Ke Liu, Bowei Gao, Liangqi Ouyang, Yingzhi Jin, Galia Pozina, Irina A. Buyanova, Weimin M. Chen, Olle Inganäs, Veaceslav Coropceanu, Jean-Luc Bredas, He Yan, Jianhui Hou, Fengling Zhang, Artem A. Bakulin, and Feng Gao. Design rules for minimizing voltage losses in high-efficiency organic solar cells. *Nature Materials*, 17:703–709, 2018.
- [131] Paul Raccuglia, Katherine C. Elbert, Philip D. F. Adler, Casey Falk, Malia B. Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A. Friedler, Joshua Schrier, and Alexander J. Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73–76, 2016.
- [132] Mahira Rafiq, Maham Salim, Safia Noreen, Rasheed Ahmad Khera, Sadia Noor, Umer Yaqoob, and Javed Iqbal. End-capped modification of dithienosilole-based small donor

- molecules for high-performance organic solar cells using a dft approach. *Journal of Molecular Liquids*, 345:117062, 2022.
- [133] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.
- [134] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: Recent applications and prospects. *npj Computational Materials*, 3(1):54, 2017.
- [135] Bharath Ramsundar. *Molecular machine learning with DeepChem*. PhD thesis, Stanford University, 2018.
- [136] Martin Riede, Danilo Spoltore, and Karl Leo. Organic solar cells-the path to commercial success. *Advanced Energy Materials*, 11:2002653, 2020.
- [137] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.
- [138] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.
- [139] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4:1256–1264, 2022.
- [140] Jerret Ross, Brian Belgodere, Samuel C. Hoffman, Vijil Chenthamarakshan, Jiri Navratil, Youssef Mroueh, and Payel Das. Gp-molformer: A foundation model for molecular generation. *Digital Discovery*, 2025. Advance Article.
- [141] Jehad Saleh, Sajjad Haider, Muhammad Saeed Akhtar, Muhammad Saqib, Muqadas Javed, Sayed Elshahat, and Ghulam Mustafa Kamal. Energy level prediction of organic semiconductors for photodetectors and mining of a photovoltaic database to search for new building units. *Molecules*, 28(3):1240, 2023.
- [142] Maham Salim, Mahira Rafiq, Rasheed Ahmad Khera, Maira Arshad, and Javed Iqbal. Amplifying the photovoltaic properties of azabodipy core based small molecules by

- terminal acceptor modification for high performance organic solar cells: A dft approach. *Solar Energy*, 233:31–45, 2022.
- [143] Benjamin Sánchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.
- [144] Narumi Sato, Sunbin Hwang, Yuichi Tsuchii, and Takuma Yasuda. Fused polycyclic lactam-based π -conjugated polymers for efficient nonfullerene organic solar cells. *Journal of Materials Chemistry A*, 11:9840–9845, 2023.
- [145] Markus C. Scharber, David Mühlbacher, Markus Koppe, Patrick Denk, Christoph Waldauf, Alan J. Heeger, and Christoph J. Brabec. Design rules for donors in bulk-heterojunction solar cells-towards 10% energy-conversion efficiency. *Advanced Materials*, 18(6):789–794, 2006.
- [146] Bettina Schweda, Matiss Reinfelds, Petra Hofstadler, Gregor Trimmel, and Thomas Rath. Recent progress in the design of fused-ring non-fullerene acceptors relations between molecular structure and optical, electronic, and photovoltaic properties. *ACS Applied Energy Materials*, 4(11):11899–11981, 2021.
- [147] Viktor Sebestyén. Renewable and sustainable energy reviews: Environmental impact networks of renewable energy power plants. *Renewable and Sustainable Energy Reviews*, 151:111626, 2021.
- [148] Martin Seifrid, Stanley Lo, Dylan G. Choi, Gary Tom, My Linh Le, Kunyu Li, Rahul Sankar, Hoai-Thanh Vuong, Hiba Wakidi, Ahra Yi, Ziyue Zhu, Nora Schopp, Aaron Peng, Benjamin R. Luginbuhl, Thuc-Quyen Nguyen, and Alan Aspuru-Guzik. Beyond molecular structure: Critically assessing machine learning for designing organic photovoltaic materials and devices. *Journal of Materials Chemistry A*, 12:14540–14558, 2024.
- [149] Xueliang Shi, Lijian Zuo, Sae Byeok Jo, Ke Gao, Francis Lin, Feng Liu, and Alex K.-Y. Jen. Design of a highly crystalline low-band gap fused-ring electron acceptor for high-efficiency solar cells with low energy loss. *Chemistry of Materials*, 29(19):8369–8376, 2017.
- [150] Yanan Shi, Yilin Chang, Kun Lu, Zhihao Chen, Jianqi Zhang, Yangjun Yan, Dingding Qiu, Yanan Liu, Muhammad Abdullah Adil, Wei Ma, Xiaotao Hao, Lingyun Zhu, and

- Zhixiang Wei. Small reorganization energy acceptors enable low energy losses in non-fullerene organic solar cells. *Nature Communications*, 13(1):1–10, 2022.
- [151] Hamed Shirzad, Ameya Velingker, Balaji Venkatachalam, Danica J Sutherland, and Ali Kemal Sinop. Exphormer: Sparse transformers for graphs. In *International Conference on Machine Learning*, 2023.
- [152] Surajudeen Sikiru, Temidayo Lekan Oladosu, Temitope Ibrahim Amosa, Sanusi Yekinni Kolawole, and Hassan Soleimani. Recent advances and impact of phase change materials on solar energy: A comprehensive review. *Journal of Energy Storage*, 53:105200, 2022.
- [153] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.
- [154] Ebru Kondolot Solak and Erdal Irmak. Advances in organic photovoltaic cells: A comprehensive review of materials, technologies, and performance. *RSC Advances*, 13(18):12244–12269, 2023.
- [155] Jinyu Sun, Dongxu Li, Jie Zou, Shaofeng Zhu, Cong Xu, Yingping Zou, Zhimin Zhang, and Hongmei Lu. Accelerating the discovery of acceptor materials for organic solar cells by deep learning. *npj Computational Materials*, 10, 2024.
- [156] Mengying Sun, Jing Xing, Huijun Wang, Bin Chen, and Jiayu Zhou. Moel: Data-driven molecular fingerprint via knowledge-aware contrastive learning from molecular graph. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3585–3594, 2021.
- [157] Rui Sun, Tao Wang, Xinrong Yang, Yao Wu, Yang Wang, Qiang Wu, Maojie Zhang, Christoph J. Brabec, Yongfang Li, and Jie Min. High-speed sequential deposition of photoactive layers for organic solar cell manufacturing. *Nature Energy*, 7:1087–1099, 2022.
- [158] Rui Sun, Yao Wu, Xinrong Yang, Yuan Gao, Zeng Chen, Kai Li, Jiawei Qiao, Tao Wang, Jing Guo, Chao Liu, Xiaotao Hao, Haiming Zhu, and Jie Min. Single-junction organic solar cells with 19.17% efficiency enabled by introducing one asymmetric guest acceptor. *Advanced Materials*, 34(26), 2022.

- [159] Wenbo Sun, Meng Li, Yong Li, Zhou Wu, Yuyang Sun, Shirong Lu, Zeyun Xiao, Baomin Zhao, and Kuan Sun. The use of deep learning to fast evaluate organic photovoltaic materials. *Advanced Theory and Simulations*, 2(1):1800116, 2019.
- [160] Wenbo Sun, Yujie Zheng, Ke Yang, Qi Zhang, Akeel A. Shah, Zhou Wu, Yuyang Sun, Liang Feng, Dongyang Chen, Zeyun Xiao, Shirong Lu, Yong Li, and Kuan Sun. Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science Advances*, 5:eaay4275, 2019.
- [161] Rakesh Suthar, T. Abhijith, and Supravat Karak. Machine-learning-guided prediction of photovoltaic performance of non-fullerene organic solar cells using novel molecular and structural descriptors. *Journal of Materials Chemistry A*, 11:22248–22258, 2023.
- [162] Nathan J. Szymanski, Christopher J. Bartel, Yan Zeng, Qingsong Tu, and Gerbrand Ceder. A probabilistic deep learning approach to automate the interpretation of multi-phase diffraction spectra. *arXiv preprint arXiv:2103.16664*, 2021.
- [163] Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624:86–92, 2023.
- [164] Ching W. Tang. Two-layer organic photovoltaic cell. *Applied Physics Letters*, 48(2):183–185, 1986.
- [165] Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark B. Gerstein. Mollm: A unified language model for integrating biomedical text with 2d and 3d molecular representations. *Bioinformatics*, 40(Suppl 1):i357–i368, 2024.
- [166] Walia Binte Tarique and Ashraf Uddin. A review of progress and challenges in the research developments on organic solar cells. *Materials Science in Semiconductor Processing*, 163, 2023.
- [167] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [168] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

- [169] Pragya Verma and Donald G. Truhlar. Status and challenges of density functional theory. *Trends in Chemistry*, 2(4):302–318, 2020.
- [170] Diego Vizoso and Remi Dingreville. Decoding diffraction and spectroscopy data with machine learning: A tutorial. *Journal of Applied Physics*, 137:131101, 2025.
- [171] Christoph Waldauf, Markus C. Scharber, Peter Schilinsky, Jens A. Hauch, and Christoph J. Brabec. Physics of organic bulk heterojunction devices for photovoltaic applications. *Journal of Applied Physics*, 99(10):104503, 2006.
- [172] Jin-Liang Wang, Fei Xiao, Jun Yan, Zhuo Wu, Kai-Kai Liu, Zheng-Feng Chang, Ru-Bo Zhang, Hui Chen, Hong-Bin Wu, and Yong Cao. Difluorobenzothiadiazole-based small-molecule organic solar cells with 8.7% efficiency by tuning of π -conjugated spacers and solvent vapor annealing. *Advanced Functional Materials*, 26(11):1803–1812, 2016.
- [173] Jingwen Wang, Huifeng Yao, Ye Xu, Lijiao Ma, and Jianhui Hou. Recent progress in reducing voltage loss in organic photovoltaic cells. *Materials Chemistry Frontiers*, 5(2):709–722, 2021.
- [174] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: Large-scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB)*, pages 429–436. ACM, 2019.
- [175] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, pages 1–9, 2022.
- [176] Zhengyang Wang, Meng Liu, Youzhi Luo, Zhao Xu, Yaochen Xie, Limei Wang, Lei Cai, Qi Qi, Zhuoning Yuan, Tianbao Yang, and Shuiwang Ji. Advanced graph and sequence neural networks for molecular property prediction and drug discovery. *Bioinformatics*, 38:2579–2586, 2022.
- [177] Yanan Wei, Zhihao Chen, Guanyu Lu, Na Yu, Congqi Li, Jinhua Gao, Xiaobin Gu, Xiaotao Hao, Guanghao Lu, Zheng Tang, Jianqi Zhang, Zhixiang Wei, Xin Zhang, and Hui Huang. Binary organic solar cells breaking 19% via manipulating the vertical component distribution. *Advanced Materials*, 34(33), 2022.

- [178] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.
- [179] Jingnan Wu, Guangwei Li, Jin Fang, Xia Guo, Lei Zhu, Bing Guo, Yulong Wang, Guangye Zhang, Lingeswaran Arunagiri, Feng Liu, He Yan, Maojie Zhang, and Yongfang Li. Random terpolymer based on thiophene-thiazolothiazole unit enabling efficient non-fullerene organic solar cells. *Nature Communications*, 11:4612, 2020.
- [180] Yao Wu, Jie Guo, Rui Sun, and Jie Min. Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *npj Computational Materials*, 6(1), 2020.
- [181] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.
- [182] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14):145301, 2018.
- [183] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 62(16):8749–8760, 2019.
- [184] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.
- [185] Xiaopeng Xu, Liyang Yu, Huifeng Meng, Liming Dai, He Yan, Ruipeng Li, and Qiang Peng. Polymer solar cells with 18.74% efficiency: from bulk heterojunction to interdigitated bulk heterojunction. *Advanced Functional Materials*, 32(4):2108797, 2022.
- [186] Zhaoyang Yao, Xunfan Liao, Ke Gao, Francis Lin, Xiaobao Xu, Xueliang Shi, Lijian Zuo, Feng Liu, Yiwang Chen, and Alex K.-Y. Jen. Dithienopicenocarbazole-based acceptors for efficient organic solar cells with optoelectronic response over 1000 nm and an extremely low energy loss. *Journal of the American Chemical Society*, 140(6):2054–2057, 2018.

- [187] Long Ye, Huawei Hu, Masoud Ghasemi, Tonghui Wang, Brian A. Collins, Joo-Hyun Kim, Kui Jiang, Joshua H. Carpenter, Hong Li, Zhengke Li, Terry McAfee, Jingbo Zhao, Xiankai Chen, Joshua Lin Yuk Lai, Tingxuan Ma, Jean-Luc Brédas, He Yan, and Harald Ade. Quantitative relations between interaction parameter, miscibility and function in organic solar cells. *Nature Materials*, 17:253–260, 2018.
- [188] Jicheng Yi, Guangye Zhang, Han Yu, and He Yan. Advantages, challenges and molecular design of different material types for solution-processed organic solar cells. *Nature Reviews Materials*, 9, 2024.
- [189] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, pages 28877–28888, 2021.
- [190] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31, 2018.
- [191] Gang Yu, Jian Gao, Jan C. Hummelen, Fred Wudl, and Alan J. Heeger. Polymer photovoltaic cells: Enhanced efficiencies via a network of internal donor-acceptor heterojunctions. *Science*, 270(5243):1789–1791, 1995.
- [192] Zhaoning Yu and Hongyang Gao. Molecular representation learning via heterogeneous motif graph neural networks. In *International Conference on Machine Learning*, pages 25581–25594. PMLR, 2022.
- [193] Jun Yuan, Tianyi Huang, Pei Cheng, Yingping Zou, Huotian Zhang, Jonathan Lee Yang, Sheng-Yung Chang, Zhenzhen Zhang, Wenchao Huang, Rui Wang, Dong Meng, Feng Gao, and Yang Yang. Enabling low voltage losses and high photocurrent in fullerene-free organic photovoltaics. *Nature Communications*, 10(1):570, 2019.
- [194] Jun Yuan, Yunqiang Zhang, Liuyang Zhou, Guichuan Zhang, Hin-Lap Yip, Tsz-Ki Lau, Xinhui Lu, Can Zhu, Hongjian Peng, Paul A. Johnson, Mario Leclerc, Yong Cao, Jacek Ulanski, Yongfang Li, and Yingping Zou. Single-junction organic solar cell with over 15% efficiency using fused-ring acceptor with electron-deficient core. *Joule*, 3(4):1140–1151, 2019.

- [195] Likai Yuan, Shijie Liang, Chengyi Xiao, Qiaomei Chen, and Weiwei Li. Near-infrared nonfullerene acceptors based on 4h-cyclopenta[1,2-b:5,4-b']dithiophene for organic solar cells and organic field-effect transistors. *Chemistry-An Asian Journal*, 16(24):4171–4178, 2021.
- [196] Xuan Zang, Xianbing Zhao, and Buzhou Tang. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry*, 6(1):34, 2023.
- [197] Ben Zhang, Fu Yang, and Yaowen Li. Recent progress in large-area organic solar cells. *Small Science*, 3(4):2300004, 2023.
- [198] Guichuan Zhang, Xian-Kai Chen, Jingyang Xiao, Philip C. Y. Chow, Minrun Ren, Grit Kupgan, Xuechen Jiao, Christopher C. S. Chan, Xiaoyan Du, Ruoxi Xia, Ziming Chen, Jun Yuan, Yunqiang Zhang, Shoufeng Zhang, Yidan Liu, Yingping Zou, He Yan, Kam Sing Wong, Veaceslav Coropceanu, Ning Li, Christoph J. Brabec, Jean-Luc Brédas, Hin-Lap Yip, and Yong Cao. Delocalization of exciton and electron wavefunction in non-fullerene acceptor molecules enables efficient organic solar cells. *Nature Communications*, 11(1), 2020.
- [199] Guichuan Zhang, Francis R. Lin, Feng Qi, Thomas Heumüller, Andreas Distler, Hans-Joachim Egelhaaf, Ning Li, Philip C. Y. Chow, Christoph J. Brabec, Alex K.-Y. Jen, and Hin-Lap Yip. Renewed prospects for organic photovoltaics. *Chemical Reviews*, 122(18):14180–14274, 2022.
- [200] Jianquan Zhang, Huei Shuan Tan, Xugang Guo, Antonio Facchetti, and He Yan. Material insights and challenges for non-fullerene organic solar cells based on small molecular acceptors. *Nature Energy*, 3:720–731, 2018.
- [201] Jiyun Zhang, Jianchang Wu, Vincent M. Le Corre, Jens A. Hauch, Yicheng Zhao, and Christoph J. Brabec. Advancing perovskite photovoltaic technology through machine learning-driven automation. *InfoMat*, 7:e70005, 2025.
- [202] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [203] Qi Zhang, Yujie Zheng, Wenbo Sun, Zeping Ou, Omololu Odunmbaku, Meng Li, Shan-shan Chen, Yongli Zhou, Jing Li, Bo Qin, and Kuan Sun. High-efficiency non-fullerene

- acceptors developed by machine learning and quantum chemistry. *Advanced Science*, 9(6), 2022.
- [204] Ting Zhang, Joshua Yuk Lin Lai, Mingzhe Shi, Qing Li, Chen Zhang, and He Yan. Data cleansing and sub-unit-based molecular description enable accurate prediction of the energy levels of non-fullerene acceptors used in organic solar cells. *Advanced Science*, 11:e2308652, 2024.
- [205] Ting Zhang, Kangzhong Wang, Kunlei Jing, Gang Li, Qing Li, Chen Zhang, and He Yan. A ring2vec description method enables accurate predictions of molecular properties in organic solar cells. *npj Computational Materials*, 10, 2024.
- [206] Wenlin Zhang, Yurong Zou, Xin Wang, Junxian Chen, and Dingguo Xu. Deep learning accelerated high-throughput screening of organic solar cells. *Journal of Materials Chemistry C*, 13:5295–5306, 2025.
- [207] Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882, 2021.
- [208] Jingbo Zhao, Yunke Li, Guofang Yang, Kui Jiang, Haoran Lin, Harald Ade, Wei Ma, and He Yan. Efficient organic solar cells processed from hydrocarbon solvents. *Nature Energy*, 1, 2016.
- [209] Can Zhu, Jun Yuan, Fangfang Cai, Lei Meng, Huotian Zhang, Honggang Chen, Jing Li, Beibei Qiu, Hongjian Peng, Shanshan Chen, Yunbin Hu, Changduk Yang, Feng Gao, Yingping Zou, and Yongfang Li. Tuning the electron-deficient core of a non-fullerene acceptor to achieve over 17% efficiency in a single-junction organic solar cell. *Energy & Environmental Science*, 13(8):2459–2466, 2020.
- [210] Jinhua Zhu, Kehan Wu, Bohan Wang, Yingce Xia, Shufang Xie, Qi Meng, Lijun Wu, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. *O*-GNN: Incorporating ring priors into molecular modeling. In *International Conference on Learning Representations (ICLR)*, 2023.
- [211] Lei Zhu, Ming Zhang, Jinqiu Xu, Chao Li, Jun Yan, Guanqing Zhou, Wenkai Zhong, Tianyu Hao, Jiali Song, Xiaonan Xue, Zichun Zhou, Rui Zeng, Haiming Zhu, Chun-Chao Chen, Roderick C. I. MacKenzie, Yecheng Zou, Jenny Nelson, Yongming Zhang, Yanming Sun, and Feng Liu. Single-junction organic solar cells with over 19% efficiency

- enabled by a refined double-fibril network morphology. *Nature Materials*, 21(6):656–663, 2022.
- [212] Federico Zipoli, Marvin Alberts, and Teodoro Laino. Ir-nmr multimodal computational spectra dataset for 177k patent-extracted organic molecules. *Scientific Data*, 12:1375, 2025.
- [213] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP*, 2010.