GLaD: Synergizing Molecular Graphs and Language Descriptors for Enhanced Power Conversion Efficiency Prediction in Organic Photovoltaic Devices

Thao Nguyen
thaotn2@illinois.edu
Siebel School of Computing and Data Science, UIUC
Champaign, Illinois, USA

Tiara Torres-Flores
Department of Chemical & Biomolecular Engineering, UIUC
Champaign, Illinois, USA
tiaract2@illinois.edu

Changhyun Hwang
Department of Chemical & Biomolecular Engineering, UIUC
Champaign, Illinois, USA
chwang12@illinois.edu

Carl Edwards
Siebel School of Computing and Data Science, UIUC
Champaign, Illinois, USA
cne2@illinois.edu

Ying Diao
Department of Chemical & Biomolecular Engineering, UIUC
Champaign, Illinois, USA
yingdiao@illinois.edu

Heng Ji
Siebel School of Computing and Data Science, UIUC
Champaign, Illinois, USA
hengji@illinois.edu

Abstract

This paper presents a novel approach for predicting Power Conversion Efficiency (PCE) of Organic Photovoltaic (OPV) devices, called **GLaD**: synergizing molecular Graphs and Language Descriptors for enhanced PCE prediction. Due to the lack of high-quality experimental data, we collect a dataset consisting of 500 pairs of OPV donor and acceptor molecules along with their corresponding PCE values, which we utilize as the training data for our predictive model. In this low-data regime, **GLaD** leverages properties learned from large language models (LLMs) pretrained on extensive scientific literature to enrich molecular structural representations, allowing for a multimodal representation of molecules. **GLaD** achieves precise predictions of PCE, thereby facilitating the synthesis of new OPV molecules with improved efficiency. Furthermore, **GLaD** showcases versatility, as it applies to a range of molecular property prediction tasks (BBBP, BACE, ClinTox and SIDER [45]), not limited to those concerning OPV materials. Especially, **GLaD** proves valuable for tasks in low-data regimes within the chemical space, as it enriches molecular representations by incorporating molecular property descriptions learned from large-scale pretraining. This capability is significant in real-world scientific endeavors like drug and material discovery, where access to comprehensive data is crucial for informed decision-making and efficient exploration of the chemical space.

CCS Concepts

• **Computing methodologies** → **Feature selection**; • **Applied computing** → **Chemistry**.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3680103

Keywords

Organic Photovoltaics, Power Conversion Efficiency Prediction, Graph Neural Network, Large Language Models

ACM Reference Format:
Thao Nguyen, Tiara Torres-Flores, Changhyun Hwang, Carl Edwards, Ying Diao, and Heng Ji. 2024. GLaD: Synergizing Molecular Graphs and Language Descriptors for Enhanced Power Conversion Efficiency Prediction in Organic Photovoltaic Devices. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3627673.3680103

1 Introduction

In materials science, the design of novel materials for organic solar cells (OSCs) is a vibrant area of research, as OSCs offer advantages such as being low-cost, flexible, and lightweight; unfortunately, they also suffer from drawbacks such as limited lifespan and poor stability [1, 6, 24]. Addressing these drawbacks necessitates the optimization of materials for OSCs, which requires quick and accurate prediction of Power Conversion Efficiency (PCE) in OSC devices to assess the quality of new candidates.

Various machine learning algorithms have been used to predict PCE of OPV devices using different datasets. Notably, the Harvard Clean Energy Project Database (CEPDB) [20] and the Harvard Organic Photovoltaic Dataset (HOPV) [31] are among the most significant public datasets in this domain. Previous studies have primarily utilized CEPDB, which comprises 2.3 million donor molecules and their corresponding PCE values calculated using Scharber's model. While training with computationally derived PCEs offers the advantage of large, standardized datasets with controlled parameters, these values often poorly correlate with experimental measurements, diminishing their practicality [16]. The HOPV dataset contains experimental PCE data for 350 different OPV donors that have been collected from various studies in the literature by Lopez et al [31], yet it lacks data on newer OPV molecules introduced after 2015, a period during which significant advancements in OPV technology achieved PCE values of up to 20% [8, 11, 14, 18]. Therefore, to expedite the development of cutting-edge OPV materials, this

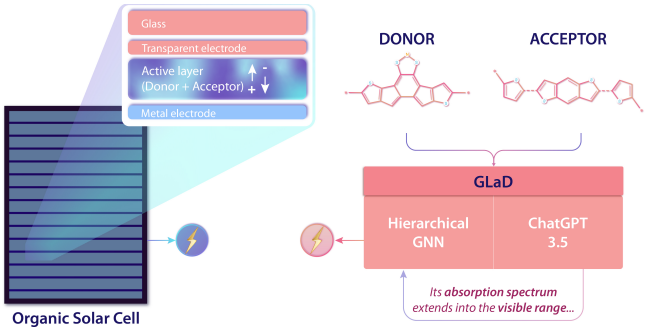

Figure 1: Overview of the GLaD PCE prediction framework.

Table 1: Statistics of our collected dataset

#samples (D-A pairs)		All 500	Train 400	Val 50	Test 50
#molecules	All	403	338	61	58
	Donor	203	165	36	32
	Acceptor	252	212	31	34
#functional modules	All	250	231	68	68
	Donor	149	136	38	38
	Acceptor	192	174	43	43
Tanimoto distance	All	0.67	0.67	0.69	0.67
	Donor	0.65	0.64	0.65	0.69
	Acceptor	0.59	0.59	0.64	0.63
PCE range		[2.5, 19.6]	[2.5, 19.6]	[6.0, 18.69]	[5.1, 17.1]

study concentrates on predicting the experimental PCE values of recently developed OPV devices only.

In this work, we present a novel approach, named **GLaD**, for accurately predicting PCE of OPV devices based on pairs of donor and acceptor molecules. To achieve this, we collected a dataset comprising 500 pairs of donors and acceptors from the literature for training our models.

GLaD addresses a key challenge in predicting PCE of OPV devices: the need for a comprehensive understanding of molecular function-structure relationships. To tackle this, chemists typically focus on the functional modules of a molecule and rely on supplementary sources like textbooks for a more comprehensive understanding of the molecule's properties. Inspired by these insights, we decompose molecules into their functional modules and integrate structural descriptors extracted by a Graph Neural Network (GNN) with textual descriptions generated by LLMs trained on extensive scientific literature. This approach aims to provide a comprehensive representation of the functional modules. After acquiring the structural and textual descriptors of the those modules, we fuse them to form a multimodal representation. Subsequently, representations of functional modules are fed into a molecule-level GNN model to predict PCE. Figure 1 illustrates the overview of **GLaD** in the PCE prediction task.

We assessed the performance of **GLaD** using our collected dataset, HOPV and the MoleculeNet benchmark [45]. Our results demonstrate that **GLaD** accurately predicts PCE values for OPV devices. Notably, incorporating textual descriptors alongside structural descriptors enhances the model's performance, with the coefficient of determination (R^2) score increasing by 0.103 ($\pm$ 0.04) in our collected dataset. For HOPV dataset, we obtain an R^2 score improvement of 0.135 compared to the baseline [10], showcasing state-of-the-art performance on this dataset. Furthermore, **GLaD** exhibits high accuracy in predicting molecular properties across various tasks (such as BBBP, BACE, ClinTox and SIDER [45]), suggesting its applicability beyond OPV-related tasks.

Our contributions are summarized as follows:

- We curate an up-to-date dataset comprising 500 pairs of donor and acceptor molecules for PCE prediction task.
- We develop a novel method, **GLaD**, that leverages learned knowledge from pretrained LLMs to generate textual descriptions for functional modules (molecular fragments) and integrates them with structural descriptors to enrich molecule representation. This approach accurately predicts PCE, achieving high R^2 scores in both our dataset and the HOPV dataset.
- **GLaD** is the first model to use a hierarchical GNN approach to integrate textual descriptions of molecular fragments (functional modules) rather than entire molecule descriptions. This improves the robustness and flexibility of our approach on unknown molecules. We conducted a study of language model-generated textual descriptions and found 88% of them to be accurate when evaluated by PhD-level domain experts.
- Our method exhibits promising results in other molecular property prediction tasks, indicating its broad applicability beyond PCE prediction.

2 Background

OSCs represent a promising class of emerging photovoltaic technologies that generate electricity from sunlight using multiple layers. In these cells, excitons (hole-electron pairs) are produced at the interface of the active layer, typically composed of a donor-acceptor (D-A) material blend of carbon-based molecules or polymers. The donor material absorbs light and generates excitons, while the acceptor material facilitates their dissociation into free charge carriers. D-A molecules thus play a crucial role in determining PCE of OPV devices by influencing light absorption, exciton dissociation, charge transport, and active layer morphology [50]. Optimizing molecular combinations can achieve broad absorption spectra, efficient exciton separation, balanced charge mobility, and ideal nanoscale phase separation, all of which enhance PCE.

Traditional methods for designing D-A combinations focus on a limited range of chemistries and typically require expert knowledge, restricting the potential for high-throughput screening [20]. However, computational modeling and machine learning offer a powerful alternative by accurately predicting the PCE of new D-A materials through the analysis of their molecular properties. This accelerates the discovery of high-performance OPV materials and broadens the range of structures explored for improved efficiency and stability.

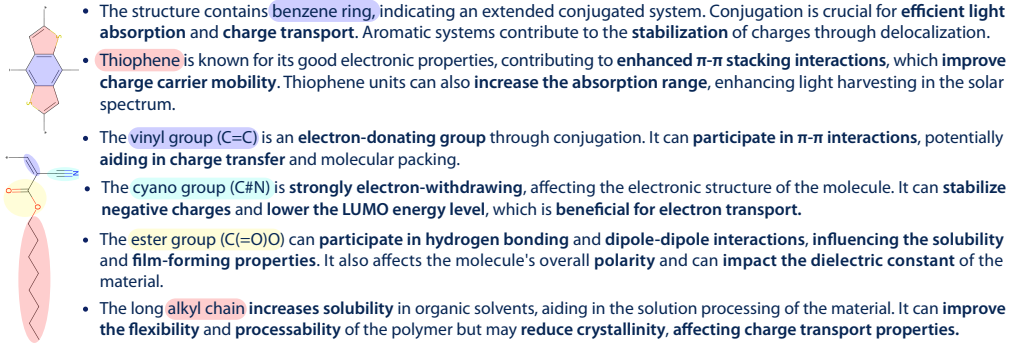

Figure 2: Examples of functional module descriptions generated by ChatGPT 3.5.

3 Related Work

3.1 PCE Prediction

The prediction of PCE has drawn considerable interest, particularly with the emergence of large datasets like CEPDB [20] and the experimental dataset HOPV [31]. Various methods are employed for this task, including quantum chemical calculations and machine learning (ML) techniques. Quantum chemical methods estimate PCE using Scharber's model [38]. This model predicts PCE of a specific OPV design based on parameters calculated by density functional theory (DFT). However, DFT calculations require significant computational time, which makes them unsuitable for quick screening [2], and there is a discrepancy between the predictions of Scharber's model and actual experimental results [16]. On the other hand, ML techniques are commonly used to explore the relationships between OPV performance and material properties more quickly and accurately [17, 44].

Many studies have focused on predicting PCE determined by Scharber's model, utilizing either the complete dataset or subsets of CEPDB [10, 23, 32, 35, 40]. Neural network architectures, including Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Graph Neural Networks (GNN), have demonstrated superior capability in learning from large datasets like CEPDB, yielding an impressive maximum R^2 score of 0.996 [10].

While some studies achieve high predictive accuracy on computational datasets, less attention is paid to the suitability of these datasets and their agreement with experimental PCE measurements. In a previous work [10], the authors investigate the impact of training data choice and conclude that while current ML models perform well on large computational datasets like CEPDB, fitting on smaller and experimental datasets proves challenging due to numerous degrees of freedom, such as experimental setups and minor device design factors. Moreover, discrepancies between computational PCE based on Scharber's model and experimental PCE are noted [16, 20], prompting efforts to collect new OPV datasets [17, 34, 42, 44]. Notably, *Greenstein et al.* constructed a new dataset comprising 1001 unique donor/non-fullerene acceptor pairs, and an ensemble of random forest and neural network models predicting PCE achieves an R^2 of 0.4 [17].

Our work extends the research on PCE prediction for OPV devices by gathering up-to-date OPV data from the literature and developing a novel predictive model. Unlike previous studies focusing solely on donor molecules [20, 31], our dataset encompasses a diverse range of donor and acceptor molecules. Each molecule is decomposed into functional modules, and an Attentive FP [46] model is employed to extract structural features from each functional module, complemented by textual descriptions generated by an LLM. This approach yields a multimodal dataset providing both structural and property knowledge of molecular functional modules, enabling precise PCE prediction with an R^2 of 0.747 (±0.04). This method also facilitates modular synthesis of new OPV molecules and sheds light on the relationship between molecular structure and PCE of OPV devices.

3.2 Multimodal Representation of Molecules: Graph Structure and Textual Descriptions

LLMs have emerged as powerful tools for molecular captioning, even from SMILES strings—compact textual representations of molecular structures [9, 19, 29]. Models like GPT (Generative Pre-trained Transformer) [36] variants can analyze these strings, generating detailed textual descriptions of molecules. Through fine-tuning on large chemical text datasets, LLMs become proficient at understanding molecular structures encoded in SMILES strings and producing coherent captions [9]. In this study, we harness the capacity of LLMs to generate structural, physical, chemical, and photovoltaic descriptions of functional modules commonly found in OPV molecules. This allows us to furnish insights into molecular properties that may not be apparent in a molecular graph without background contextual information. Additionally, this method enhances the factual correctness of the generated text, given the relative ease with which LLMs generate captions for shorter SMILES strings (molecular substructures) compared to longer ones (the entire molecule). We note that functional modules are molecular subgraphs often referred to as fragments in other work [12].

Several previous studies have focused on incorporating SMILES strings and textual descriptions to enhance molecular understanding tasks. In earlier works [30, 49], a unified representation of text and SMILES was created by replacing chemical compound names in text with SMILES strings. Other studies [7, 26, 27, 39, 51] aligned

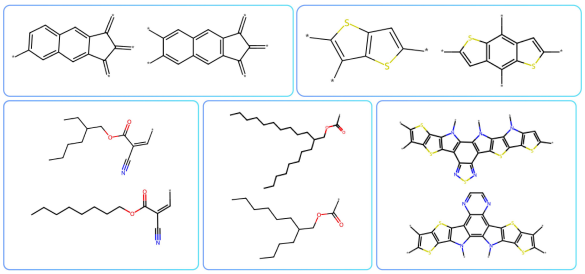

Figure 3: Pairs of fragments that are close in the text embedding space, indicating shared properties or functions. Fragments with similar structures, such as side chains containing ester groups (C(=O)O) and cyano groups (C#N), are clustered together in the embedding space, reflecting their structural similarities.

SMILES strings and textual descriptions through contrastive learning or cross-modal projection to ensure that their representations are close in the representation space. Both methods achieved high performance in molecular understanding tasks, with MolXPT [30] achieving state-of-the-art results in MoleculeNet tasks [45].

While several studies have combined knowledge graphs and text descriptions to enhance the representations of either modality or both [21], no prior research has integrated textual data into graphs of molecular fragments (functional modules). In this study, for the first time, we integrate structural embeddings obtained from a GNN model and text embeddings obtained from LLMs to form a multimodal representation of such functional modules. By doing so, our model can make predictions based on information from both modalities, ultimately enhancing its performance on a wide variety of prediction tasks.

4 OPV Dataset Collection

Due to the lack of curated high-quality experimental data, we curated an OPV dataset to train our PCE prediction model. The dataset consists of 500 pairs of donor and acceptor molecules employed in bulk heterojunction (BHJ) and bilayer OPV devices collected from literature from 2012 to 2023.

In this dataset, there are a total of 403 molecular entities (comprised by 10 atoms: C, H, O, N, S, Si, Se, Cl, Br, F), including 203 donor molecules and 252 acceptor molecules (with 52 molecules that can be either donor or acceptor in a device). It includes properties of OPV devices such as PCE, open circuit potential (V_{oc}), short circuit current density (J_{sc}), and fill factor (FF) for each donor-acceptor pair. Table 1 provides the statistics of the collected dataset. Compared to the HOPV dataset [31], our dataset demonstrates superior diversity, encompassing a significantly larger portion of the chemical space. We attribute this to five key differences:

(1) It contains pairs of donor-acceptor molecules, instead of solely donor molecules as in HOPV;
(2) It includes up-to-date data of OPV devices with a higher PCE range, from 2.5% to 19.6%, compared to 0.0005% to 10.2% in HOPV;
(3) It comprises molecules of greater diversity, reflected in a lower average Tanimoto distance [3] of 0.67, compared to 0.8 in HOPV;

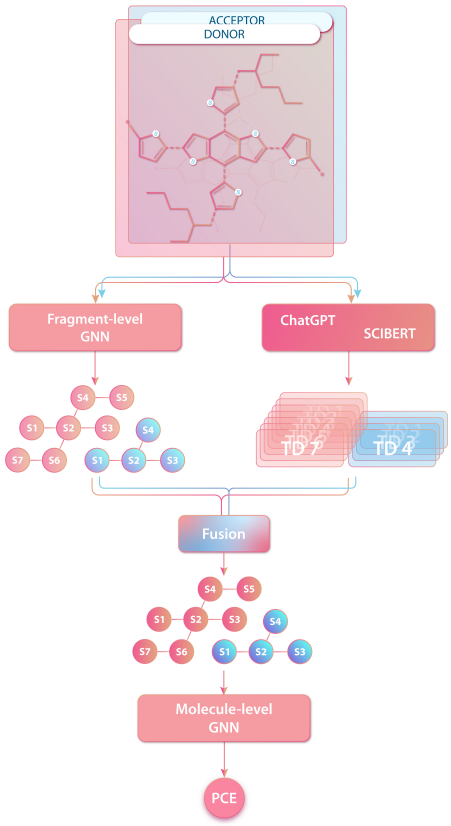

Figure 4: Proposed model architecture (TD: Textual descriptors of a functional module).

(4) It contains a more diverse range of atom types (10 atoms) compared to the 8 atom types present in HOPV (C, H, O, N, F, S, Si, and Se);
(5) It contains a larger number of samples (500 samples compared to 350 in HOPV).

Each molecule in the dataset is further decomposed into functional modules, also referred to as fragments, for additional processing. A total of 250 different functional modules result from the decomposition of the 403 molecules in the dataset.

With this dataset, our objective is to construct machine learning models capable of accurately predicting the PCE score based on pairs of donor and acceptor molecules. A robust PCE prediction model is characterized by a high coefficient of determination (R^2), low Mean Square Error (MSE), and low Mean Absolute Error (MAE).

5 Fusing Text with Molecular Structure

In this section, we detail our approach, **GLaD**, for extracting structural and textual descriptors for each functional module, and then fusing them to form the multimodal representation of those modules. Figure 4 illustrates the architecture of our proposed model.

Table 2: PCE prediction results on the collected dataset with different GNN architectures (average of 30 runs).

	Hierarchical GNN	Molecule-Level GNN	Donor-only GNN	Acceptor-only GNN
MSE	**3.58**	4.61	10.18	6.65
95% CI	[2.81, 4.41]	[3.82, 5.31]	[8.69, 11.79]	[5.67, 7.41]
MAE	**1.461**	1.792	2.726	2.162
95% CI	[1.28, 1.66]	[1.62, 1.95]	[2.45, 2.95]	[1.99, 2.33]
$\mathbf{R^2}$	**0.644**	0.534	0.398	0.428
95% CI	[0.59, 0.68]	[0.48, 0.59]	[0.34, 0.45]	[0.38, 0.48]

5.1 Modeling Molecular Structure

After collecting the SMILES strings of OPV molecules, we construct molecular graphs and employ a molecular decomposition algorithm to decompose them into constituent functional modules. This algorithm breaks down molecules at C-C single bonds between conjugated backbone rings and their corresponding side chains. This approach harnesses modular synthesis, wherein complex molecules are iteratively assembled from smaller constituent functional modules [5, 15, 25].

Fragment-level graphs representing functional modules will undergo processing by a GNN model to produce structural descriptors. Various GNN architectures, including Graph Convolutional Networks (GCN) [22], Graph Attention Networks (GAT) [41], and Attentive FP [46], are employed to extract structural descriptors from molecular graphs. Subsequently, these structural descriptors of each functional module will be fused with textual descriptors to create a multimodal representation of each functional module.

5.2 Generating Textual Descriptions for Functional Modules

For each functional module, ChatGPT-3.5 [33] is utilized to generate descriptions including their structural, physical, chemical, and photovoltaic properties. A total of 250 descriptions are produced. These descriptions then undergo manual evaluation to ensure the factual accuracy of the generated text. A subset of 60 functional modules and their descriptions is manually evaluated, revealing that 88% (53 out of 60) are correct. Figure 2 exemplifies a result generated by GPT-3.5.

To generate text descriptions for functional modules, we use their SMILES string to query ChatGPT-3.5 [33] with this prompt: *Generate descriptions of this molecular fragment:* [*SMILES*] *focusing on its structural, physical, chemical, and photovoltaic properties. Descriptions should be specific and tailored for organic photovoltaic (OPV) material research. Avoid neutral information.*

5.3 Modeling Textual Descriptions

Textual descriptions of functional modules are fed into a frozen Scibert [4] model to extract text embeddings. We assessed the efficacy of combining descriptions for each property with structural descriptors to identify those yielding improvements, which will be retained for the generation of textual descriptors.

In order to evaluate the quality of textual descriptors, we randomly selected data points that are close to each other in the text embedding space. Outcomes (depicted in Figure 3) show that functional modules with similar structures tend to cluster together in the text embedding space, suggesting that textual descriptors effectively capture information regarding the similarity of molecular fragments.

5.4 Fusion Approaches

After generating both structural and textual descriptors for functional modules, we combine them using fusion operators. We evaluate two fusion operators: *average + concat* and *attention + concat*.

The first approach computes an embedding for the entire text description by averaging all the word embeddings, and then concatenates this with the structural embedding to form a multimodal representation of the functional module, denoted as $\mathbf{v}$.

The attention-based module comprises learnable query ($\mathbf{W}_Q$), key ($\mathbf{W}_K$), and value ($\mathbf{W}_V$) matrices to learn the cross-attention score between the structural embedding vector $\mathbf{s}$ of a functional module and the word embedding vectors $\mathbf{t}$ of its description. The attention weight is calculated by Equation 1.

$$\alpha = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{1}$$

Here, α represents the cross-attention score between $\mathbf{s}$ and $\mathbf{t}$, where $\mathbf{Q} = \mathbf{W}_Q \cdot \mathbf{t}$, $\mathbf{K} = \mathbf{W}_K \cdot \mathbf{s}$, and $\mathbf{V} = \mathbf{W}_V \cdot \mathbf{s}$. The term d_k denotes the dimensionality of the key $\mathbf{s}$.

The embedding of the entire text description is computed as the weighted average of word embeddings, with the attention scores from the structural embedding serving as the weights, as shown in Equation 2.

$$\mathbf{t}' = \sum_{i=0}^{N} \alpha_i \cdot \mathbf{t}_i \tag{2}$$

Here N represents the length of the text description.

Finally, the structural and textual embeddings are concatenated to create a multimodal representation of each functional module $\mathbf{v}$, expressed as $\mathbf{v} = \text{concat}(\mathbf{s}, \mathbf{t}')$.

After fusion, each functional module is represented by a vector $\mathbf{v}$, representing a node in the molecule-level graph. The edges of this graph are defined by the bonds connecting the functional modules. This graph is input to the molecule-level GNN model, which outputs a predicted PCE score for the input donor-acceptor pair.

6 Experiments

6.1 Experimental Settings

Datasets and evaluation metrics. We conduct an evaluation of **GLaD** across multiple datasets, including our collected OPV dataset, the HOPV dataset [31], and several tasks from the MoleculeNet benchmark dataset (BBBP, BACE, ClinTox, and SIDER) [45]. To evaluate the efficacy of our proposed method across both computational and experimental data, we assess its performance on the HOPV dataset for experimental PCE and PCE computed using Scharber's model [38]. We utilize three commonly used metrics: R^2, MSE, and MAE for PCE prediction task. Meanwhile, for the MoleculeNet tasks, we employ the AUC-ROC metric to evaluate its performance.

Table 3: PCE prediction results on the collected dataset of the baseline model (hierarchical GNN without incorporating textual descriptors) and models incorporated with different kinds of textual descriptors.

		Baseline	Full description	Structural	Physical	Chemical	Photovoltaic	Physical+ Chemical
MSE	**Avg**	3.583	2.878	2.66	2.561	2.308	3.317	2.327
	95% CI	[2.81, 4.412]	[1.282, 3.51]	[1.445, 4.283]	[1.883, 3.322]	[1.851, 2.858]	[2.579, 4.176]	[1.584, 3.186]
MAE	**Avg**	1.461	1.32	1.289	1.231	1.218	1.367	1.194
	95% CI	[1.282, 1.664]	[1.174, 1.482]	[1.053, 1.563]	[1.059, 1.42]	[1.08, 1.366]	[1.171, 1.56]	[0.985, 1.479]
R^2	**Avg**	0.644	0.703	0.725	0.732	0.735	0.659	**0.747**
	(↑)	-	↑ 0.059	↑ 0.081	↑ 0.088	↑ 0.091	↑ 0.015	↑ **0.103**
	95% CI	[0.658, 0.757]	[0.688, 0.759]	[0.694, 0.774]	[0.688, 0.779]	[0.611, 0.703]	[0.703, 0.794]	[0.698, 0.787]

Table 4: PCE prediction results on the HOPV dataset of the proposed method and SVR model [10]

		w/o text	w/ text	SVR
MSE	**Avg**	2.598	**2.321**	2.687
	±	0.524	0.487	0.487
MAE	**Avg**	1.233	**1.034**	1.132
	±	0.146	0.136	0.095
R^2	**Avg**	0.492	**0.588**	0.453
	±	0.109	0.115	0.109

Table 5: Results of predicting computational PCE on the HOPV dataset, using computational PCE obtained from Scharber's model with a selection of four functionals (B3LYP, BP86, M06-2X, and PBE0).

		B3LYP		BP86		M06-2X		PBE0	
		w/o text	w/ text	w/o text	w/ text	w/o text	w/ text	w/o text	w/ text
MSE	**Avg**	0.064	0.035	3.487	0.188	2e-4	1e-4	0.036	0.003
	±	0.022	0.01	1.185	0.064	8e-5	2e-5	0.011	6e-4
MAE	**Avg**	0.182	0.136	1.354	0.273	0.031	0.005	0.133	0.038
	±	0.025	0.014	0.191	0.035	0.002	0.001	0.019	0.004
R^2	**Avg**	0.943	**0.968**	0.935	**0.964**	0.966	**0.974**	0.951	**0.996**
	±	0.02	0.019	0.018	0.02	0.019	0.021	0.017	8e-4

Table 6: Results of GLaD on four common MoleculeNet tasks compared to other GNN-based models. Evaluation metric: ROC-AUC(%)

Dataset	**BBBP**	**BACE**	**ClinTox**	**SIDER**
#molecules	2039	1513	1478	1427
#tasks	1	1	2	27
D-MPNN [47]	71.0 (0.3)	80.9 (0.6)	90.6 (0.6)	57.0 (0.7)
AttentiveFP [46]	64.3 (1.8)	78.4 (0.02)	84.7 (0.3)	60.6 (3.2)
GROVER$_{large}$ [37]	69.5 (0.1)	81.0 (1.4)	76.2 (3.7)	65.4 (0.1)
MolCLR [43]	72.2 (2.1)	82.4 (0.9)	**91.2** (3.5)	58.9 (1.4)
GraphMVP [28]	72.4 (2.1)	81.2 (0.9)	79.1 (2.8)	63.9 (1.2)
GEM [13]	72.4 (0.4)	85.6 (1.1)	90.1 (1.3)	67.2 (0.4)
HiMo$_{small}$ [48]	71.3 (0.6)	84.6 (0.2)	70.6 (2.1)	62.5 (0.3)
HiMo$_{large}$ [48]	73.2 (0.8)	84.3 (0.3)	80.8 (1.4)	61.3 (0.5)
GLaD (w/o text)	82.8 (1.2)	82.1 (0.8)	85.6 (1.7)	64.3 (0.9)
GLaD (w/ text)	**86.4** (1.5)	**85.7** (0.9)	87.3 (1.2)	**68.1** (1.3)

Data split. In accordance with the experimental setups in previous work [10], we split the two OPV datasets into training, validation, and test sets with a ratio of 80:10:10. Similarly, for the MoleculeNet tasks, we split the dataset into training, validation, and test sets with ratio of 80:10:10, respectively, following previous research [52].

GLaD's model architecture. We evaluate various model architectures by experimenting with the following setups, testing their performance on the collected dataset:

- Different GNNs (including a GNN that takes a molecular graph as input and directly outputs predictions of PCE, versus a hierarchical GNN including a fragment-level GNN that extracts structural descriptors of functional modules followed by a molecule-level GNN that takes multimodal representations of fragments as input);
- Different kinds of textual descriptions (structural, physical, chemical, photovoltaic property descriptions, and descriptions of all properties).

6.2 Main Results

Results on the collected dataset: The results of using a molecule-level GNN and a hierarchical GNN (without text) are described in Table 2. According to our findings, using a hierarchical GNN architecture that combines fragment-level and molecule-level GNNs results in a significant improvement in the R^2 score of 0.11 (± 0.04) when compared to using only molecule-level GNN. We also examine the effectiveness of using only donor or acceptor molecules as input for the hierarchical GNN model. We find that R^2 score can be greatly increased by using pairs of donor and acceptor molecules as input. This leads to an R^2 score of 0.644 (± 0.05), whereas models that use either only donor or only acceptor molecules have R^2 of 0.398 and 0.428, respectively.

Table 3 shows experimental results of incorporating various kinds of textual descriptors with structural descriptors obtain from the fragment-level GNN. We observe that using textual description of all properties improve predictive performance from 0.015 to 0.103 in R^2 score, with the highest improvement from physical and chemical descriptions, and the combination of both.

Results on the HOPV dataset. The results from experiments employing the proposed **GLaD** model on the HOPV dataset are presented in Table 4 and Table 5. These results demonstrate that **GLaD** outperforms another method using the SVR model by 0.135 in R^2 score in the task of predicting experimental PCE. Table 5 also demonstrates **GLaD**'s ability to accurately predict computational PCE, achieving the highest R^2 score of 0.996. It is worth noting that complementing structural descriptors with textual descriptors consistently improves the predictive performance of the model, both in the collected dataset and HOPV dataset.

Table 7: PCE prediction results by GLaD (with textual descriptors being physical + chemical property descriptions) using various GNN models for fragment-level and molecule-level GNNs (first column: fragment-level GNN + molecule-level GNN).

	MSE		MAE		R^2	
	Avg	Std	Avg	Std	Avg	Std
GAT + GAT	4.617	0.811	1.7	0.122	0.564	0.083
GCN + GCN	5.924	1.08	1.94	0.19	0.407	0.128
Attentive FP + GAT	4.424	0.517	1.59	0.115	0.564	0.084
GAT + Attentive FP	4.064	0.657	1.674	0.135	0.598	0.097
Attentive FP + GCN	3.481	0.922	1.457	0.181	0.648	0.091
GCN + Attentive FP	5.715	0.978	1.82	0.242	0.476	0.098
GAT + GCN	3.884	0.791	1.621	0.221	0.603	0.092
GCN + GAT	4.672	0.98	1.734	0.133	0.559	0.105
Attentive FP + Attentive FP	**2.327**	0.801	**1.194**	0.141	**0.747**	0.088

Table 8: GLaD's PCE prediction results on the collected dataset using average + concat (Avg) and attention + concat (Att) as fusion operators.

	Full description		Structural		Physical		Chemical		Photovoltaic		Physical + Chemical
	Avg	Att	Avg	Att	Avg	Att	Avg	Att	Avg	Att	Avg	Att
MSE	2.437	2.878	3.346	2.66	2.762	2.561	2.541	2.308	4.274	3.317	2.343	2.327
MAE	1.293	1.32	1.464	1.289	1.276	1.231	1.311	1.218	1.507	1.367	1.253	1.194
R^2	**0.712**	0.703	0.634	**0.725**	0.712	**0.732**	0.705	**0.735**	0.622	**0.659**	0.719	**0.747**

Results on the MoleculeNet datasets. Table 6 shows the results of using **GLaD** to solve four classification tasks with small datasets in MoleculeNet [45]. **GLaD** outperforms other GNN-based models on 3 out of 4 tasks, achieving a significant margin in the BBBP task with a gap of 13.2 (± 1.5)% compared to the second-best method (HiMo [48]). These results demonstrate **GLaD**'s ability to excel in tasks beyond PCE prediction, proving particularly valuable for low-resource tasks where data collection is challenging. This demonstrates incorporating fragment-level text descriptions can significantly enrich molecule representation.

7 Ablation Study

GNN architectures for fragment-level GNN and molecule-level GNN. We experimented with various setups for two GNN models: the fragment-level GNN and the molecule-level GNN. The results are shown in Table 7, indicating that using Attentive FP for both levels gave us the best predictions, scoring an impressive R^2 of 0.747 (± 0.04).

Fusion methods. For the fusion block, we utilize *average + concat* and *attention + concat* as fusion methods and present the results in Table 8. Experimental results show that *attention + concat* proves more robust than *average + concat* in 5 out of 6 setups. Hence, for **GLaD**, we opt for *attention + concat* over *average + concat* as the fusion method for integrating structural descriptors and textual descriptors of functional modules.

Node features for fragment-level GNN. We also evaluate the effectiveness of utilizing various atomic properties as node features for the fragment-level GNN in **GLaD**. Experimental results shown in Table 9 demonstrate that solely employing electronegativity as a node feature yields the highest R^2 score of 0.747 (± 0.04). Consequently, we opt for the electronegativity of atoms as the node attribute for **GLaD**.

Table 9: Experimental results of using different kinds of atomic properties as node features for the fragment-level GNN in GLaD, with textual descriptors being physical and chemical property descriptions.

Features	MSE		MAE		R^2	
	Avg	Std	Avg	Std	Avg	Std
Atomic number	3.343	1.03	1.353	0.159	0.679	0.107
Mass	3.513	0.948	1.467	0.173	0.664	0.108
Electronegativity (EN)	**2.327**	0.8	**1.194**	0.141	**0.747**	0.088
EN + hybridization	2.865	0.516	1.357	0.118	0.731	0.061
EN + degree	3.197	0.787	1.405	0.162	0.676	0.074
EN + formal charge	2.554	0.371	1.352	0.09	0.743	0.046
EN + implicit & explicit valance	2.951	0.532	1.337	0.129	0.7	0.069
EN + is aromatic	3.199	0.803	1.366	0.171	0.668	0.084
All	2.59	0.42	1.322	0.117	0.736	0.052

8 Conclusion & Future Work

In this study, we introduce a new dataset and present a novel approach for predicting PCE in OPV devices. Our approach leverages the learned properties of LLMs to enrich molecular representations at the level of functional modules (molecular fragments). This representation enables accurate prediction of the PCE of OPV devices, as well as other property prediction tasks. However, to apply PCE prediction to high-throughput screening of OPV materials, enhancing prediction reliability is crucial. To achieve this, we plan to incorporate uncertainty quantification methods at both the molecular and functional module levels into our future work. By doing so, we aim to further strengthen our predictions and advance the field of OPV material screening.

Ethics Considerations

In conducting this study, we ensured ethical compliance by sourcing data from publicly available scientific literature and databases, adhering to standards of data integrity and transparency. Our model, **GLaD**, is intended to complement human expertise and support decision-making in scientific research, particularly in low-data regimes. We advocate for responsible use of **GLaD** by validating its predictions with domain experts to ensure safety and reliability in applications such as drug and material discovery. Additionally, we are committed to transparency and reproducibility by sharing our methodologies and findings with the broader scientific community.

Acknowledgements

This work is supported by the Molecule Maker Lab Institute: an AI research institute program supported by NSF under award No. 2019897. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of, the National Science Foundation, and the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Omar A Abdulrazzaq, Viney Saini, Shawn Bourdo, Enkeleda Dervishi, and Alexandru S Biris. 2013. Organic solar cells: a review of materials, limitations, and possibilities for improvement. *Particulate science and technology* 31, 5 (2013), 427–442.

[2] Carlo Adamo and Denis Jacquemin. 2013. The calculations of excited-state properties with Time-Dependent Density Functional Theory. *Chemical Society Reviews* 42, 3 (2013), 845–856.

[3] Dávid Bajusz, Anita Rácz, and Károly Héberger. 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics* 7 (2015), 1–13.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

[5] Daniel J. Blair, Chitti Chitti, Sriyankari, Melanie Trobe, David M Kostyra, Hannah M. S. Haley, Richard L. Hansen, Steve G. Ballmer, Toby J. Woods, Wesley Wang, Vikram Mubayi, Michael J. Schmidt, Robert W. Pipal, Greg F. Morehouse, Andrea M. Palazzolo Ray, Danielle L. Gray, Adrian L. Gill, and Martin D. Burke. 2022. Automated iterative Csp3–C bond formation. *Nature* 604, 7904 (2022), 92–97.

[6] Quinn Burlingame, Melissa Ball, and Yueh-Lin Loo. 2020. It's time to focus on organic solar cell stability. *Nature Energy* 5, 12 (2020), 947–949.

[7] He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208* (2023).

[8] Yong Cui, Ye Xu, Huifeng Yao, Pengqing Bi, Hong Ling, Jianqi Zhang, Yunfei Zu, Tao Zhang, Jinzhao Qin, Junzhen Ren, Zhihao Chen, Chang He, Xiaotao Hao, Zhixiang Wei, and Jianhui Hou. 2021. Single-junction organic photovoltaic cell with 19% efficiency. *Advanced Materials* 33, 41 (2021), 2102420.

[9] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817* (2022).

[10] Andreas Eibeck, Daniel Nurkowski, Angiras Menon, Jiaru Bai, Jinkui Wu, Li Zhou, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. 2021. Predicting power conversion efficiency of organic photovoltaics: models and data analysis. *ACS omega* 6, 37 (2021), 23764–23775.

[11] Mohamed Boudia El Amine, Zhou Yi, Hongying Li, Qiuwang Wang, Jun Xi, and Cunlu Zhao. 2023. Latest updates of single-junction organic solar cells up to 20% efficiency. *Energies* 16 (2023), 3895.

[12] Peter Ertl and Ansgar Schuffenhauer. 2009. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics* 1 (2009), 1–11.

[13] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. 2022. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence* 4, 2 (2022), 127–134.

[14] Jiehao Fu, Patrick W. K. Fong, Heng Liu, Chieh-Szu Huang, Xinhui Lu, Shirong Lu, Maged Abdelsamie, Tim Kodalle, Carolin M. Sutter-Fella, Yang Yang, and Gang Li. 2023. 19.31% binary organic solar cell and low non-radiative recombination enabled by non-monotonic intermediate state transition. *Nature Communications* 14, 1 (2023), 1760.

[15] Eric P Gillis and Martin D Burke. 2009. Iterative cross-couplng with MIDA boronates: towards a general platform for small molecule synthesis. *Aldrichimica acta* 42, 1 (2009), 17.

[16] Brianna L Greenstein, Danielle C Hiener, and Geoffrey R Hutchison. 2022. Computational evolution of high-performing unfused non-fullerene acceptors for organic solar cells. *The Journal of Chemical Physics* 156, 17 (2022).

[17] Brianna L Greenstein and Geoffrey R Hutchison. 2023. Screening efficient tandem organic solar cells with machine learning and genetic algorithms. *The Journal of Physical Chemistry C* 127, 13 (2023), 6179–6191.

[18] Shitao Guan, Yaokai Li, Chang Xu, Chenran Xu, Mengting Wang, Yuxi Xu, Qi Chen, Dawei Wang, Lijian Zuo, and Hongzheng Chen. 2024. Self-assembled interlayer enables high-performance organic photovotlaics with power conversion efficiency exceeding 20%. *Advanced Materials* 2400342 (2024), 2400342.

[19] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems* 36 (2023), 59662–59688.

[20] 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. 2011. 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 (2011), 2241–2251.

[21] Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2023. Large language models on graphs: A comprehensive survey. *arXiv preprint arXiv:2312.02783* (2023).

[22] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[23] Xiangyu Kong and Bohao Xu. 2023. Prediction of photoelectric conversion efficiency of organic photovoltaic materials based on deep learning. In *Ninth International Conference on Mechanical Engineering, Materials, and Automation Technology (MMEAT 2023)*, Vol. 12801. SPIE, 248–253.

[24] Pankaj Kumar and Suresh Chand. 2012. Recent progress and future aspects of organic solar cells. *Progress in Photovoltaics: Research and applications* 20, 4 (2012), 377–415.

[25] Junqi Li, Anthony S Grillo, and Martin D Burke. 2015. From synthesis to function via iterative assembly of N-methyliminodiacetic acid boronate building blocks. *Accounts of chemical research* 48, 8 (2015), 2297–2307.

[26] Pengfei Liu, Yiming Ren, and Zhixiang Ren. 2023. GIT-Mol: A Multi-modal Large Language Model for Molecular Science with Graph. *Image, and Text* (2023).

[27] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence* 5, 12 (2023), 1447–1457.

[28] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2021. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728* (2021).

[29] Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798* (2023).

[30] Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023. Molxpt: Wrapping molecules with text for generative pre-training. *arXiv preprint arXiv:2305.10688* (2023).

[31] Steven A Lopez, Edward O Pyzer-Knapp, Gregor N Simm, Trevor Lutzow, Kewei Li, Laszlo R Seress, Johannes Hachmann, and Alán Aspuru-Guzik. 2016. The Harvard organic photovoltaic dataset. *Scientific data* 3, 1 (2016), 1–7.

[32] Steven A Lopez, Benjamin Sanchez-Lengeling, Julio de Goes Soares, and Alán Aspuru-Guzik. 2017. Design principles and top non-fullerene acceptor candidates for organic photovoltaics. *Joule* 1, 4 (2017), 857–870.

[33] OpenAI. 2021. ChatGPT 3.5. https://openai.com/gpt-3/. Accessed: January 2022.

[34] Daniele Padula, Jack D Simpson, and Alessandro Troisi. 2019. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Materials Horizons* 6, 2 (2019), 343–349.

[35] Edward O Pyzer-Knapp, Kewei Li, and Alan Aspuru-Guzik. 2015. Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery. *Advanced Functional Materials* 25, 41 (2015), 6495–6502.

[36] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[37] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems* 33 (2020), 12559–12571.

[38] Markus C Scharber, David Mühlbacher, Markus Koppe, Patrick Denk, Christoph Waldauf, Alan J Heeger, and Christoph J Brabec. 2006. Design rules for donors in bulk-heterojunction solar cells—Towards 10% energy-conversion efficiency. *Advanced materials* 18, 6 (2006), 789–794.

[39] Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481* (2022).

[40] Wenbo Sun, Meng Li, Yong Li, Zhou Wu, Yuyang Sun, Shirong Lu, Zeyun Xiao, Baomin Zhao, and Kuan Sun. 2019. The use of deep learning to fast evaluate organic photovoltaic materials. *Advanced Theory and Simulations* 2, 1 (2019), 1800116.

[41] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[42] Hongshuai Wang, Jie Feng, Zhihao Dong, Lujie Jin, Miaomiao Li, Jianyu Yuan, and Youyong Li. 2023. Efficient screening framework for organic solar cells with deep learning and ensemble learning. *npj Computational Materials* 9, 1 (2023), 200.

[43] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* 4, 3 (2022), 279–287.

[44] Yao Wu, Jie Guo, Rui Sun, and Jie Min. 2020. Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *npj Computational Materials* 6, 1 (2020), 120.

[45] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* 9, 2 (2018), 513–530.

[46] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. 2019. Pushing the boundaries of molecular representation for drug discovery with

the graph attention mechanism. *Journal of medicinal chemistry* 63, 16 (2019), 8749–8760.

[47] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* 59, 8 (2019), 3370–3388.

[48] Xuan Zang, Xianbing Zhao, and Buzhou Tang. 2023. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry* 6, 1 (2023), 34.

[49] Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications* 13, 1 (2022), 862.

[50] Min Zhang, Lei Zhu, Guanqing Zhou, Tianyu Hao, Chaoqun Qiu, Zhe Zhao, Qin Hu, Bryon W. Larson, Haiming Zhu, Zaifei Ma, Zheng Tang, Wei Feng, Yongming Zhang, Thomas P. Russell, and Feng Liu. 2021. Single-layered organic photovoltaics with double cascading charge transport pathways: 18% efficiencies. *Nature Communications* 12, 309 (2021).

[51] Haiteng Zhao, Shengchao Liu, Ma Chang, Hannan Xu, Jie Fu, Zhihong Deng, Lingpeng Kong, and Qi Liu. 2024. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *Advances in Neural Information Processing Systems* 36 (2024).

[52] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: A universal 3d molecular representation learning framework. (2023).