Molecular Hypergraph Neural Networks

Junwu Chen[1,2] and Philippe Schwaller[1,2]
[1]*Laboratory of Artificial Chemical Intelligence (LIAC), Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.*
[2]*National Centre of Competence in Research (NCCR) Catalysis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.*

(*Electronic mail: philippe.schwaller@epfl.ch)

(Dated: 22 December 2023)

Graph neural networks (GNNs) have demonstrated promising performance across various chemistry-related tasks. However, conventional graphs only model the pairwise connectivity in molecules, failing to adequately represent higher-order connections like multi-center bonds and conjugated structures. To tackle this challenge, we introduce molecular hypergraphs and propose Molecular Hypergraph Neural Networks (MHNN) to predict the optoelectronic properties of organic semiconductors, where hyperedges represent conjugated structures. A general algorithm is designed for irregular high-order connections, which can efficiently operate on molecular hypergraphs with hyperedges of various orders. The results show that MHNN outperforms all baseline models on most tasks of OPV, OCELOTv1 and PCQM4Mv2 datasets. Notably, MHNN achieves this without any 3D geometric information, surpassing the baseline model that utilizes atom positions. Moreover, MHNN achieves better performance than pretrained GNNs under limited training data, underscoring its excellent data efficiency. This work provides a new strategy for more general molecular representations and property prediction tasks related to high-order connections.

I. INTRODUCTION

Graph presentation of molecular structures, also called molecular graphs, finds extensive application in computational chemistry and machine learning, where atoms are served as nodes and chemical bonds as edges. Graph neural networks (GNNs) are a class of deep learning models that can handle graph-structured data and are related to geometric deep learning[1–5]. Unlike traditional neural networks that operate on regular grids (e.g., images) or sequential data (e.g., text), GNNs can handle interconnected and non-Euclidean data, making them suitable for tasks involving graphs with complex topologies[4]. This inherent advantage enables GNNs to directly learn the complex topological relationships of atoms and chemical bonds through molecular graphs[6]. In recent years, GNNs have demonstrated excellent molecular representation capabilities and achieved promising performance on many chemistry-related tasks, such as molecular property prediction[6–8], drug design[9–11], interatomic potentials[12–14], spectroscopic analysis[15–17], reaction prediction and retrosynthesis[18–20].

However, ordinary graphs are limited to modeling pairwise connectivity within molecular structures, falling short in effectively representing higher-order connections[11,21,22]. A substantial number of molecules have delocalized bonds, such as multi-center bonds[23] and conjugated bonds[24]. In contrast to classical chemical bonds localized between pairs of atoms, each delocalized bond involves three or more atoms[25]. As illustrated in Figure 1a, two B atoms and one H atom share two electrons to form a 3-center-2-electron bond, which cannot be represented by a pairwise edge[26]. Similarly, conjugated organic molecules like porphyrin in Figure 1b, possess long-range dispersed π electrons beyond the descriptive capability of conventional edges[24]. Therefore, the development of a more comprehensive graph representation for molecular structures becomes imperative to address this limitation inherent to conventional graphs.

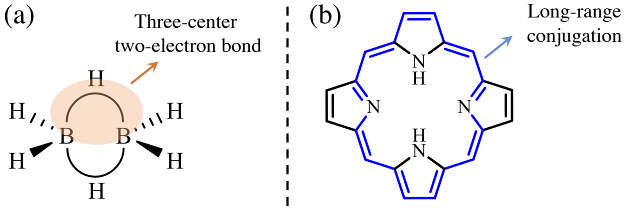

FIG. 1. (a) Diborane structure and its 3-center-2-electron bond (B-H-B). (b) Porphyrin structure and its long-range conjugated bond.

A hypergraph is a generalization of the graph where a hyperedge can join any number of nodes[27,28]. Due to the innate ability to capture higher-order relationships, hypergraphs can powerfully model complex topological structures such as social networks[29], chemical reactions[30], and compound–protein interactions[11,31,32]. Hypergraph Neural Networks (HGNs) belong to a category of neural networks designed to work with hypergraphs and extend the idea of GNNs to handle hyperedges[28,31]. Several studies[33,34] have employed HGNs in the field of chemistry and depicted atoms as hyperedges and bonds between two atoms as nodes. While these approaches improve the validity of molecule generation and enhance edge representation learning[33,34], they presently do not leverage hyperedges to articulate high-order connections within molecules. For diverse molecular structures, especially organometallic complexes, and conjugated molecules, hyperedges from hypergraphs are competent to represent multi-atomic connections like delocalized bonds due to their inherent advantages[35,36].

Conjugated molecules, characterized by alternating single and multiple bonds along a molecular backbone, play a pivotal role in photoelectric applications such as organic light-emitting diodes (OLEDs) and organic solar cells (OSCs)[37,38]. Their distinctive advantage stems from the delocalized π electrons

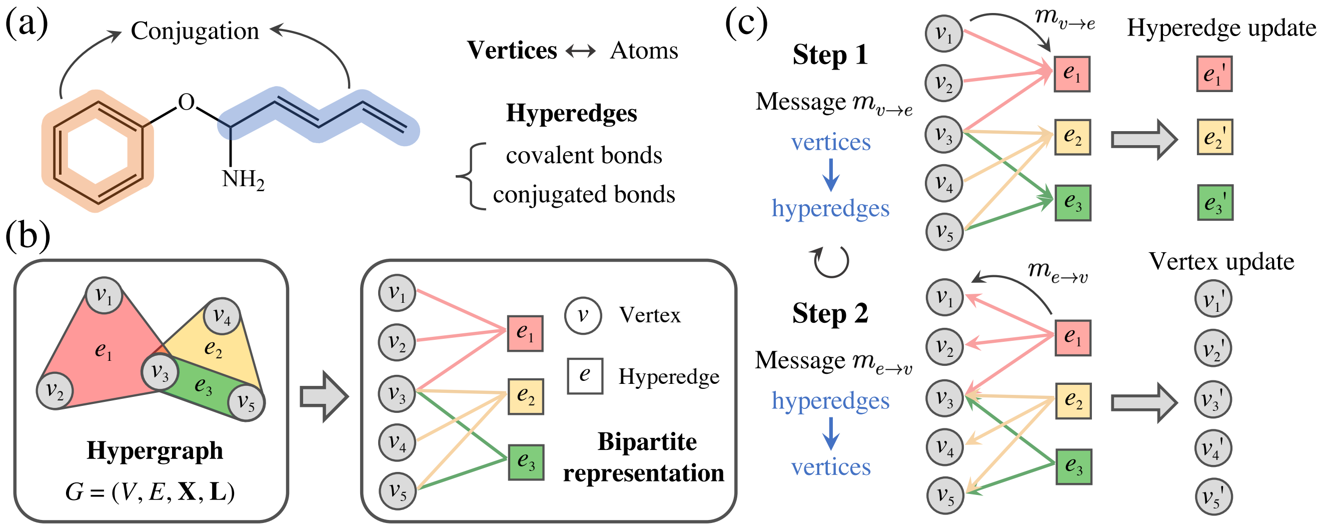

FIG. 2. (a) The method of constructing molecular hypergraphs for conjugated molecules. (b) The conversion from a hypergraph to an equivalent bipartite graph. (c) The message passing method of our MHNN model.

within conjugated structures, which can facilitate charge transport and optical absorption, establishing them as indispensable components of organic semiconductors[38]. Although various machine learning models, especially GNNs, have been developed for predicting optoelectronic properties and accelerating the design of organic semiconductors[39–42], high-order conjugated connections have still not been properly modeled.

Herein, we introduce the concept of molecular hypergraphs and propose a Molecular Hypergraph Neural Network (MHNN) based on a simple but general message-passing method. MHNN was implemented to predict the optoelectronic properties of organic semiconductors where hyperedges represent conjugated structures. On three photovoltaic-related datasets, MHNN outperforms all baseline models in most tasks. Despite not using any 3D geometric information, MHNN exhibits better results than 3D-based models like SchNet[43] which require atom coordinates as input. Moreover, MHNN possesses high data efficiency even compared with pretrained models, which could be useful for data-scarce applications. This work provides a new model for property prediction of complex molecules containing higher-order connections.

II. METHODS

A. Molecular hypergraph

A hypergraph $G = (V, E, \mathbf{H}, \mathbf{L})$ is defined by a set of n nodes V, a set of m hyperedges E, node features $\mathbf{H} \in \mathbb{R}^{n\times d}$, and hyperedge features $\mathbf{L} \in \mathbb{R}^{m\times d'}$. Each hyperedge $e = \left\{v_1, \cdots, v_{|e|}\right\}$ is a subset of V and its order $|e| \geq 2$. In a molecular hypergraph, it is natural to employ nodes to represent atoms and hyperedges to represent pairwise bonds, delocalized bonds, conjugated bonds and other higher-order associations. It is worth noting that the definition of hyperedges is important and should be related to the prediction target. For example, conjugated structures can significantly affect the light absorption and emission of molecules, so it is reasonable to describe conjugated bonds with hyperedges for the prediction of optoelectronic properties (e.g., bandgap).[38] Moreover, hyperedges could be defined by pharmacophores[44] or toxicophores[45] for the prediction of molecular activity or toxicity, respectively. In this work, we show an example of using molecular hypergraphs to describe conjugated molecules (Fig. 2a), where hyperedges are constructed by pairwise bonds and conjugated bonds. Like benzene (C_6H_6) containing 12 atoms, six C-H σ bonds, six C-C σ bonds, and one large delocalized π bond, its molecular hypergraph consists of twelve nodes, twelve 2-order hyperedges, and one 6-order hyperedge.

B. Algorithm

The higher-order relations in complex molecules are often very diverse, that is, the orders of hyperedges in molecular hypergraphs often vary. For example, the number of atoms contained in a conjugated bond can be any integer greater than four. Therefore, model algorithms should not be limited to hyperedges of a specific order or within a specific order range. In addition, the model should also have good extrapolation ability for hyperedges of unseen orders. Inspired by recent works about hypergraph diffusion algorithms[46,47], we propose the Molecular Hypergraph Neural Networks (MHNN) based on bipartite representations of hypergraphs, which can efficiently operate on hypergraphs with hyperedges of various orders (Fig. 2bc).

The molecular hypergraph is initially transformed into an equivalent bipartite graph (Fig. 2b), wherein two distinct sets of vertices denote the nodes and hyperedges of the molecular hypergraph, respectively. The message passing of MHNN

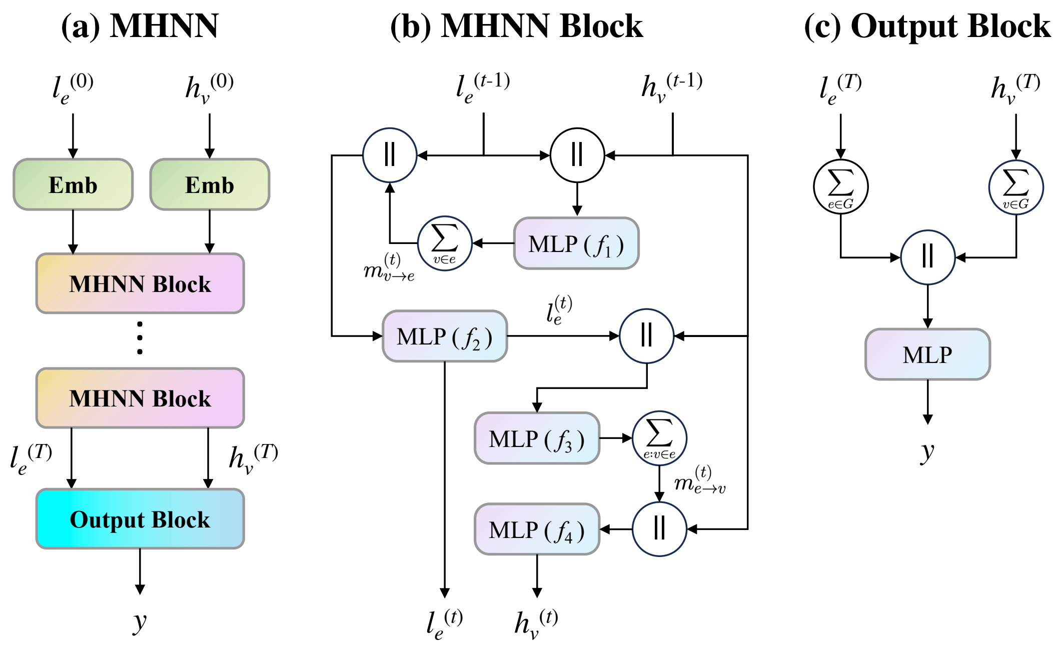

FIG. 3. The MHNN architecture. || denotes concatenation. The embeddings of nodes and hyperedges are updated in multiple MHNN blocks which can share parameters or not. The final embeddings of nodes and hyperedges are passed into an output block to generate predictions.

relies on the bipartite representations converted from molecular hypergraphs. Each message passing layer of MHNN is defined in terms of four differentiable functions f_1, f_2, f_3, and f_4. In the t $(1 \le t \le T)$ step message passing, the hidden states $l_e^{(t)}$ of each hyperedge are updated based on the messages $m_{v\to e}^{(t)}$ from the connected nodes ($v \in e$) according to:

$$m_{v\to e}^{(t)} = \sum_{v\in e} f_1\left(h_v^{(t-1)}, l_e^{(t-1)}\right) \tag{1}$$

$$l_e^{(t)} = f_2\left(l_e^{(t-1)}, m_{v\to e}^{(t)}\right) \tag{2}$$

Then, the hidden states $h_v^{(t)}$ of each node are updated based on the messages $m_{e\to v}^{(t)}$ from involved hyperedges ($e : v \in e$) according to:

$$m_{e\to v}^{(t)} = \sum_{e:v\in e} f_3\left(l_e^{(t)}, h_v^{(t-1)}\right) \tag{3}$$

$$h_v^{(t)} = f_4\left(h_v^{(t-1)}, m_{e\to v}^{(t)}\right) \tag{4}$$

where $h_v^{(0)}$ and $l_e^{(0)}$ are derived from initial atom features and bond features (Appendix B). After T steps message passing, the hypergraph-level prediction is calculated in the readout part based on the final hidden states of nodes and hyperedges ($|e| > 2$), according to:

$$\hat{y} = \mathrm{MLP}\left(\sum_{v\in G} h_v^{(T)}, \sum_{e\in G} l_e^{(T)}\right) \tag{5}$$

where MLP$(\cdot)$ is a Multi-Layer Perceptron. The output $\hat{y}$ is the prediction target of MHNN, which can be a scalar or a vector. In this work, four MLPs are used to act as update functions (f_1, f_2, f_3, f_4). The schematic diagram of MHNN architecture is shown in Fig. 3 and Algorithm 1.

Algorithm1 Algorithm of MHNN

```
Input: molecular hypergraph $G = (V, E, \mathbf{H}, \mathbf{L})$
Initialization: four MLPs $(f_1, f_2, f_3, f_4)$ in each MHNN block, which can share parameters across $T$ layers or not. One MLP in the output block.
for $t = 1, 2, ..., T$ do
    Send messages from $V$ to $E$ for all $e \in E$:
    $m_{v\to e}^{(t)} = \sum_{v\in e} f_1\left(\left[h_v^{(t-1)}, l_e^{(t-1)}\right]\right)$
    Update hyperedge embeddings $l_e^{(t)} = f_2\left(\left[l_e^{(t-1)}, m_{v\to e}^{(t)}\right]\right)$
    Send messages from $E$ to $V$: $m_{e\to v}^{(t)} = \sum_{e:v\in e} f_3\left(\left[l_e^{(t)}, h_v^{(t-1)}\right]\right)$
    Update node embeddings $h_v^{(t)} = f_4\left(\left[h_v^{(t-1)}, m_{e\to v}^{(t)}\right]\right)$
end for
hypergraph embedding from nodes: $g_v = \sum_{v\in G} h_v^{(T)}$
hypergraph embedding from hyperedges: $g_e = \sum_{e\in G} l_e^{(T)}$, $|e| > 2$
$\hat{y} = \mathrm{MLP}([g_v, g_e])$
Output: $\hat{y}$
```

C. Input features

For 2D GNN baselines, the atoms features and bond features designed by OGB[48] are used for the initial features of models. For MHNN, initial atom features are from OGB[48] and only bond types are used as the initial feature of all hyperedges. For

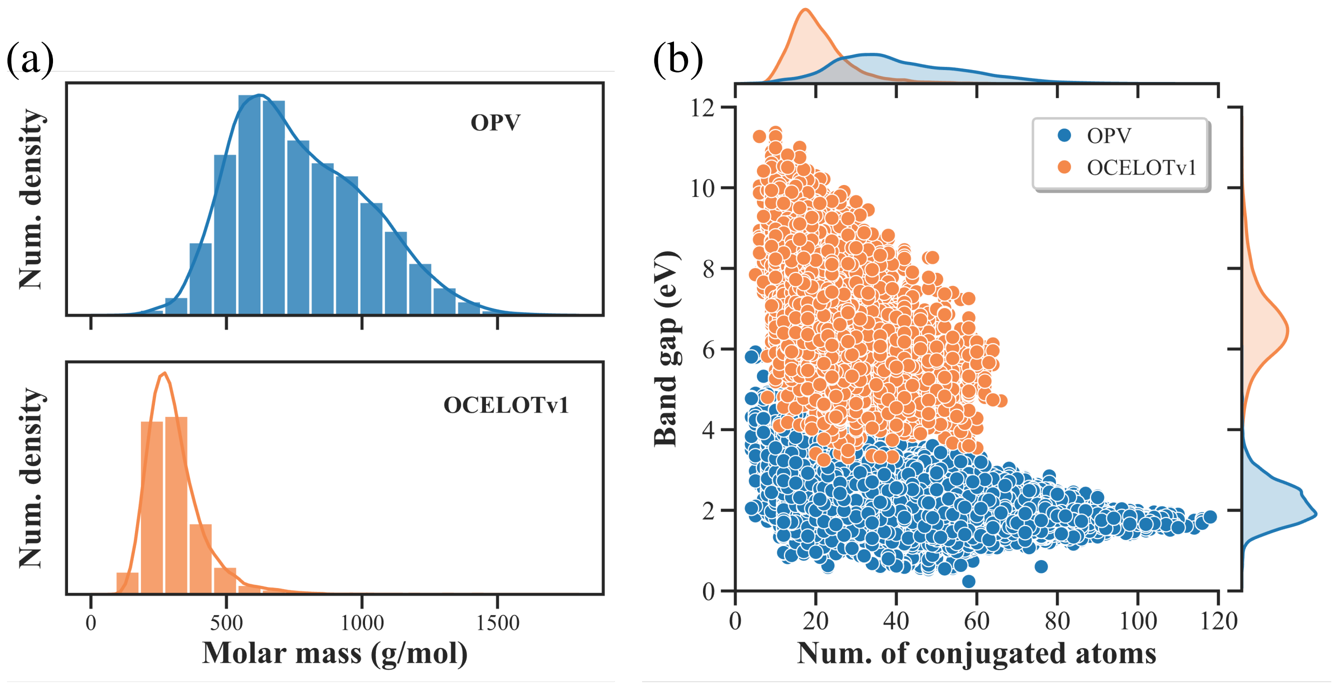

FIG. 4. (a) Distribution of molecular weights for OPV and OCELOTv1 datasets. (b) Distribution of band gap and atomic number of conjugated structures for OPV and OCELOTv1 datasets.

3D GNN baselines, only atomic numbers are used as the initial node feature. More details are listed in the Appendix B.

D. Datasets

TABLE I. Overview of the datasets

Dataset	Graphs	Task type	Task number	Metric
OPV	90,823	regression	8	MAE
OCELOTv1	25,251	regression	15	MAE
PCQM4Mv2	3,746,620	regression	1	MAE

The OPV dataset[39], named organic photovoltaic dataset, contains 90,823 unique molecules (monomers and soluble small molecules) and their SMILES strings, 3D geometries, and optoelectronic properties from DFT calculations. OPV has four molecular tasks for monomers, the energy of highest occupied molecular orbital ($\varepsilon_{\mathrm{HOMO}}$), lowest unoccupied molecular orbital ($\varepsilon_{\mathrm{LUMO}}$), HOMO-LUMO gap ($\Delta\varepsilon$), and the spectral overlap $I_{overlap}$. In addition, OPV has four polymeric tasks, the polymer $\varepsilon_{\mathrm{HOMO}}$, polymer $\varepsilon_{\mathrm{LUMO}}$, polymer gap $\Delta\varepsilon$, and optical LUMO O_{LUMO}.[39]

The OCELOTv1 dataset[40] comprises about 25,000 organic π-conjugated molecules, along with their optoelectronic and reaction characteristics calculated by precise DFT or TD-DFT methods. The dataset encompasses 15 molecular properties: vertical (VIE) and adiabatic (AIE) ionization energy, vertical (VEA) and adiabatic (AEA) electron affinity, cation (CR) and anion (AR) relaxation energy, HOMO and LUMO energy, HOMO–LUMO energy gap (H–L), electron (ER) and hole (HR) reorganization energy, and lowest-lying singlet (S0S1) and triplet (S0T1) excitation energy.

PCQM4Mv2[49] is based on the PubChemQC project[50] and aims to predict the HOMO-LUMO energy gap of molecules from SMILES strings. PCQM4Mv2 is unprecedentedly large (> 3.8M graphs) compared to other labeled graph-related databases.

We follow the standard train/validation/test dataset splits from OPV and PCQM4Mv2, and use random split for the OCELOT dataset. The experimental results are derived from three separate runs using different random seeds, except for PCQM4Mv2, which is based on one single random seed run.

III. RESULTS AND DISCUSSION

In this section, we initially assessed the predictive performance of MHNN on optoelectronic properties across three datasets. Among them, the OPV[39] and OCELOTv1[39] datasets consist of conjugated molecules and their optoelectronic properties, while the PCQM4Mv2 dataset was employed to investigate the large-scale learning capability of MHNN. Subsequently, we explored the data efficiency of MHNN at different training data sizes.

A. Analysis of datasets

OPV and OCELOTv1 datasets, composed of conjugated molecules, are utilized to explore the learning ability of MHNN

TABLE II. MAE results on OPV testing set. The unit of $I_{overlap}$ target is W/mol, and the unit of other targets is meV. * represents using DFT-optimized atom coordinates during model training. The results of MPNN and SchNet are from the reference[39].

Methods	Molecular				Polymer			
	$\Delta\varepsilon$	$\varepsilon_{\mathrm{HOMO}}$	$\varepsilon_{\mathrm{LUMO}}$	$I_{overlap}$	$\Delta\varepsilon$	$\varepsilon_{\mathrm{HOMO}}$	$\varepsilon_{\mathrm{LUMO}}$	O_{LUMO}
GCN	67.9 ± 1.2	38.2 ± 0.3	55.3 ± 1.5	265.8 ± 4.4	76.2 ± 1.4	54.2 ± 0.5	61.8 ± 0.6	61.6 ± 0.5
GIN	48.5 ± 0.4	29.2 ± 0.2	38.6 ± 0.6	188.8 ± 2.8	66.8 ± 0.7	48.8 ± 0.4	54.9 ± 0.6	54.0 ± 0.3
GAT	54.7 ± 1.2	33.5 ± 0.7	42.9 ± 1.6	204.2 ± 7.7	72.5 ± 1.7	51.6 ± 1.2	58.6 ± 0.8	58.0 ± 0.6
GATv2	57.7 ± 2.3	32.6 ± 1.5	44.0 ± 2.2	200.1 ± 1.3	73.1 ± 0.8	51.9 ± 0.3	57.8 ± 0.8	58.2 ± 0.8
MPNN	36.9 ± 0.4	32.1 ± 0.8	27.9 ± 0.7	149.3 ± 2.3	57.1 ± 0.5	49.1 ± 0.8	47.8 ± 0.7	47.8 ± 0.5
SchNet*	32.7 ± 0.5	27.0 ± 0.4	24.8 ± 0.4	**96.6 ± 0.9**	69.8 ± 0.6	56.9 ± 0.3	56.8 ± 0.5	57.2 ± 0.3
MHNN	**28.6 ± 0.2**	**22.1 ± 0.1**	**21.2 ± 0.3**	113.5 ± 0.7	**56.6 ± 0.1**	**45.8 ± 0.7**	**45.1 ± 0.3**	**44.7 ± 0.1**

on conjugated structure and its prediction performance for optoelectronic properties. As shown in Fig. 4a, the conjugated molecules in the OPV dataset have a broader molar mass distribution (80-1800 g/mol) compared to the OCELOTv1 dataset (90-1400 g/mol). The molecular weights in the OPV dataset are predominantly concentrated in the range of 500 to 1000, whereas the OCELOTv1 dataset shows a concentration in the range of 200 to 400. Therefore, the OPV dataset not only has more data points than the OCELOTv1 dataset, but also has more large conjugated molecules. As depicted in Fig. 4b, molecules with larger conjugated structures are present in the OPV dataset compared to the OCELOTv1 dataset. The number of atoms in each conjugated structure of the OPV dataset spans a range from 4 to 120, with a concentration between 25 and 50. In contrast, the OCELOTv1 dataset exhibits a narrower range of atom numbers of conjugated structures (5-66), and is mainly concentrated between 15 and 30. Moreover, the conjugated molecules in the OPV dataset generally have lower band gaps ($\sim$ 1.9 eV) compared to the OCELOTv1 dataset ($\sim$ 6.2 eV). It can be concluded from Fig. 4b that molecules with larger conjugated structures tend to have smaller band gaps, but this is not absolute. The distribution without obvious regularity also demonstrates the complex relationship between the photoelectric properties and conjugated structures. This underscores the significance of utilizing hyperedges to represent conjugated structures.

B. Performance on OPV dataset

For OPV dataset, we compared MHNN with multiple baselines: GCN[51], GIN[52], GAT[53], GATv2[54], MPNN[55] and SchNet[43]. Table II shows the test performances of MHNN and competitive baselines on the OPV dataset, where the best results are marked in bold. Except for SchNet[43] which uses the 3D molecular geometries from DFT calculations, other models including MHNN, only use 2D topology information from SMILES strings. As for molecular properties, SchNet is obviously better than the 2D baselines, since 3D information is crucial for these properties[39]. However, MHNN outperforms all baselines on three tasks ($\Delta\varepsilon$, $\varepsilon_{\mathrm{HOMO}}$, $\varepsilon_{\mathrm{LUMO}}$) without any 3D information, indicating that molecular hypergraphs with additional conjugation information are reliable representations of organic semiconductors. The SchNet model outperforms other models significantly in the prediction of the target $I_{overlap}$, indicating that the 3D molecular geometries can provide crucial and unique insights for predicting this target. For polymer property prediction tasks, SchNet[43] cannot exhibit better performance because only atom positions of monomers are available. It also suggests that polymer properties could be less dependent on the precise 3D structures of monomers[39]. Overall, MHNN achieves the best results on 7 out of 8 tasks compared to baselines, which demonstrates the significance of molecular hypergraphs and the excellent performance of MHNN for property prediction of conjugated molecules.

C. Performance on OCELOTv1 dataset

All models from the original paper[40] were selected as baseline models to compare the performance of MHNN on the OCELOTv1 dataset. Extended connectivity fingerprint (ECFP2) and 266 molecular descriptors were calculated from SMILES strings and used as the input for ridge regression (RR), support vector machine (SVM), kernel ridge regression (KRR) and feed-forward network (FFN)[40]. For the MPNN+MolDes model, the graph embeddings computed by MPNN are concatenated with the vectors of molecular descriptors, and employed for predicting molecular properties through a FFN[40]. More details about the baseline models can be found in Reference[40]. Table III shows the test performances of MHNN and baselines, where the best results are marked in bold. On the tasks such as AIE, AEA, S0S1 and S0T1, MPNN exhibits better performance than models (RR, SVM, KRR, FFN) using molecular descriptors. However, the models using molecular descriptors show superior performance than MPNN in the tasks like HOMO, H-L and HR. Moreover, with the assistance of extra molecular descriptors, MPNN+MolDes model demonstrates greater predictive performance across most tasks compared to other models. It indicates that both molecular graphs and molecular descriptors can provide important and specific information for the optoelectronic property prediction, respectively. Despite not using molecular descriptors, MHNN outperforms all baseline models in 15 tasks, demonstrating its excellent prediction performance. This illustrates that molecular hypergraphs are strong representations of conjugated molecules and

TABLE III. MAE results of baselines and MHNN on OCELOTv1 testing set. The unit of all targets is eV. The results of baselines are from the reference[40].

Target	RR	SVM	KRR	FFN	MPNN	MPNN+MolDes	**MHNN**
HOMO	0.345 ± 0.005	0.317 ± 0.003	0.337 ± 0.003	0.354 ± 0.012	0.796 ± 0.446	0.330 ± 0.028	**0.306 ± 0.004**
LUMO	0.340 ± 0.006	0.277 ± 0.005	0.306 ± 0.002	0.297 ± 0.004	0.291 ± 0.044	0.289 ± 0.028	**0.258 ± 0.003**
H-L	0.580 ± 0.005	0.604 ± 0.006	0.561 ± 0.004	0.578 ± 0.011	1.264 ± 0.696	0.548 ± 0.029	**0.519 ± 0.011**
VIE	0.231 ± 0.004	0.204 ± 0.002	0.241 ± 0.004	0.219 ± 0.001	0.202 ± 0.043	0.191 ± 0.024	**0.178 ± 0.003**
AIE	0.222 ± 0.002	0.193 ± 0.002	0.222 ± 0.004	0.207 ± 0.003	0.176 ± 0.015	0.173 ± 0.006	**0.162 ± 0.004**
CR1	0.058 ± 0.001	0.059 ± 0.001	0.057 ± 0.001	0.063 ± 0.001	0.054 ± 0.001	0.055 ± 0.002	**0.053 ± 0.001**
CR2	0.059 ± 0.001	0.061 ± 0.001	0.056 ± 0.001	0.059 ± 0.001	0.061 ± 0.001	0.053 ± 0.001	**0.052 ± 0.000**
HR	0.112 ± 0.001	0.114 ± 0.001	0.113 ± 0.001	0.110 ± 0.002	0.126 ± 0.022	0.133 ± 0.019	**0.099 ± 0.001**
VEA	0.218 ± 0.004	0.172 ± 0.002	0.231 ± 0.004	0.186 ± 0.002	0.193 ± 0.052	0.157 ± 0.018	**0.138 ± 0.001**
AEA	0.210 ± 0.001	0.182 ± 0.002	0.219 ± 0.002	0.176 ± 0.002	0.160 ± 0.027	0.154 ± 0.027	**0.124 ± 0.002**
AR1	0.057 ± 0.001	0.053 ± 0.001	0.057 ± 0.001	0.062 ± 0.002	0.057 ± 0.002	0.051 ± 0.001	**0.050 ± 0.001**
AR2	0.052 ± 0.001	0.051 ± 0.001	0.053 ± 0.000	0.051 ± 0.001	0.048 ± 0.002	0.052 ± 0.001	**0.046 ± 0.001**
ER	0.104 ± 0.020	0.099 ± 0.002	0.105 ± 0.002	0.101 ± 0.002	0.093 ± 0.002	0.098 ± 0.006	**0.092 ± 0.001**
S0S1	0.307 ± 0.006	0.275 ± 0.004	0.307 ± 0.002	0.282 ± 0.003	0.252 ± 0.017	0.249 ± 0.013	**0.241 ± 0.003**
S0T1	0.230 ± 0.003	0.183 ± 0.003	0.235 ± 0.004	0.194 ± 0.003	0.148 ± 0.012	0.150 ± 0.028	**0.145 ± 0.002**

MHNN can extract important information related to optoelectronic properties from conjugated structures.

D. Performance on PCQM4Mv2 dataset

TABLE IV. Validate MAE results of MHNN and other message-passing GNN baselines on the PCQM4Mv2. The results of baselines are from the reference[49,56]. This dataset does not publish its test set. VN represents the use of virtual nodes to improve performance.

Model	Parameters	Validate MAE (eV)
GCN	2.0 M	0.1379
GIN	3.8 M	0.1195
GAT	6.7 M	0.1302
GCN-VN	4.9 M	0.1153
GAT-VN	6.7 M	0.1192
MHNN	2.1 M	**0.1125**

To explore the learning ability on large-scale dataset, MHNN is compared with GNN baselines with a message passing mechanism on the PCQM4Mv2 dataset (Table IV). It should be pointed out that there are a large number of small molecules without conjugated structures in this dataset, even though the prediction target is band gap, one of the optoelectronic properties. As shown in Table IV, MHNN can obtain lower MAE results with fewer model parameters, which proves its high learning efficiency. This also shows that MHNN has reliable large-scale learning ability and could reduce the training cost on huge datasets.

E. Data efficiency

To explore the data efficiency of MHNN, we compare it to GIN with or without pretraining on the three most important tasks of OPV dataset under the same data partition. All 80,823 unlabelled molecules in the training set were used to pretrain the GIN model using self-supervised learning (SSL) strategy[57]. Different amounts of data were randomly selected from the training set to directly train GIN and MHNN or finetune the pretrained GIN. As shown in Figure 5, MHNN exhibits better results on three tasks than GIN and pretrained GIN at the different training data sizes. For instance, using 1000 labeled training data, MHNN surpasses pretrained GIN by 31% and 25% on the $\varepsilon_{\mathrm{HOMO}}$ and $\varepsilon_{\mathrm{LUMO}}$ tasks, respectively. In addition, directly-trained GIN needs 4∼6 times more training data to attain performance equivalent to MHNN. All the results show that MHNN is highly data-efficient and could be useful for applications without abundant labeled data.

IV. CONCLUSION

The molecular hypergraph and corresponding MHNN were designed to overcome the limitations of traditional molecular graphs when it comes to representing high-order connections within complex molecules. The photoelectric property prediction task of organic semiconductors was selected to evaluate its prediction performance. The definition of molecular hyperedges is specified to focus on conjugated structures of molecules, which relies on human knowledge of relevant connections rather than learning directly from data. Across all three datasets (OPV, OCELOTv1, PCQM4Mv2), MHNN exhibits superior performance to the baselines on most tasks. Impressively, even in the absence of 3D geometric information, MHNN surpasses SchNet which relies on atom positions. Moreover, MHNN demonstrates higher data efficiency compared to pretrained models, making it valuable for applications where labeled data is scarce.

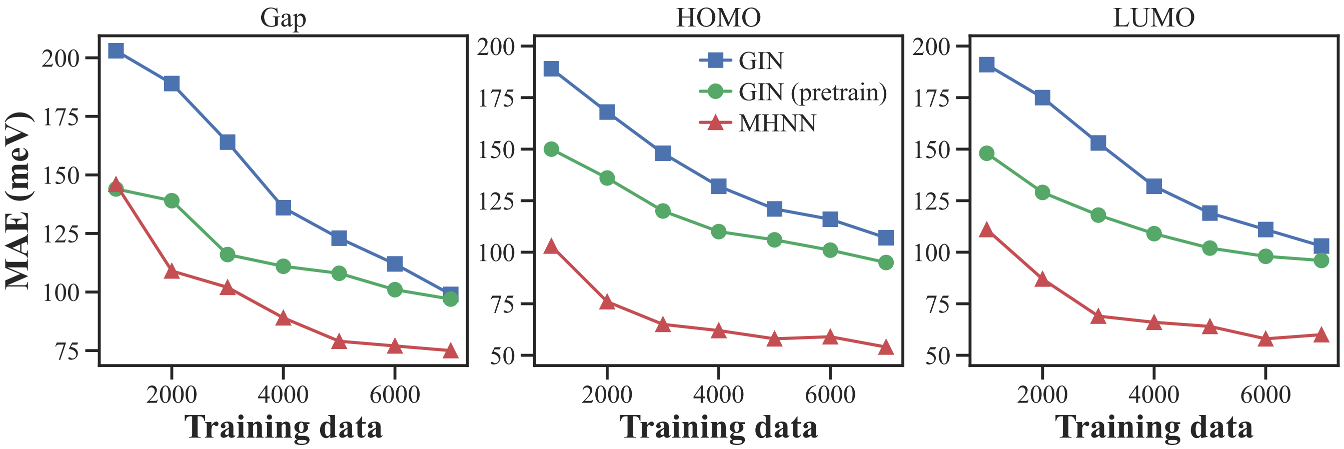

FIG. 5. The test results of different models on the HOMO-LUMO gap, HOMO and LUMO tasks of OPV dataset under different amounts of training data. The green lines represent the results of pretrained GIN by self-supervised learning[57], while the blue and red lines show the results from GIN and MHNN without pretraining, respectively. Except for the MHNN model, all data are from the reference[57].

ACKNOWLEDGMENTS

This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

CODE AVAILABILITY

The Python code of MHNN and baseline models for optoelectronic property prediction on the OPV, OCELOTv1, and PCQM4Mv2 dataset can be found on GitHub at https://github.com/schwallergroup/mhnn.

Appendix A: Implementation Details

Our implementation is based on PyTorch and PyG[58,59]. The code of 2D GNN baselines is from OGB[48]. The experiments were conducted in a collaborative computing cluster setting, featuring diverse CPU and GPU architectures. This included a combination of NVidia V100 (32GB) and RTX3090 (24GB) GPUs. For a fair comparison, the same training recipe was used for all the models on the same dataset. For baseline models, the hyperparameters were adopted from references[39,49].

Appendix B: Input features

The Tables B.1, B.2, and B.3 describe the input features for atoms, pair-wise edges, and hyperedges.

TABLE B.1. Atom (node) features for MHNN and 2D GNN baselines.

Feature	Description
Atom type	type of atom (ex. C, N, O), by atomic number
Chirality	unspecified, tetrahedral CW/CCW, or other
Degree	number of bonds the atom is involved in
Formal charge	integer electronic charge assigned to atom
Hydrogens	number of bonded hydrogen atoms
Radical electrons	the number of unpaired electrons
Hybridization	sp, sp2, sp3, sp3d, or sp3d2
Aromaticity	whether this atom is part of an aromatic system
Is in ring	whether the atom is in a ring

TABLE B.2. Bond (edge) features for 2D GNN baselines.

Feature	Description
Bond type	single, double, triple, or aromatic
Bond stereo	none, any, E/Z or cis/trans
Is conjugated	whether the bond is conjugated

TABLE B.3. Using bond type as the hyperedge feature of MHNN.

Edge order	Feature
$=2$	bond type: single, double, triple, or aromatic
>2	conjugated bonds

REFERENCES

[1] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," Advances in neural information processing systems **28** (2015).

[2] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning* (PMLR, 2017) pp. 1263–1272.

[3] J. Gasteiger, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," arXiv preprint arXiv:2003.03123 (2020).

[4] P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer, *et al.*, "Graph neural networks for materials science and chemistry," Communications Materials **3**, 93 (2022).

[5] K. Atz, F. Grisoni, and G. Schneider, "Geometric deep learning on molecular representations," Nature Machine Intelligence **3**, 1023–1032 (2021).

[6] X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu, and H. Wang, "Geometry-enhanced molecular representation learning for property prediction," Nature Machine Intelligence **4**, 127–134 (2022).

[7] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, *et al.*, "Analyzing learned molecular representations for property prediction," Journal of chemical information and modeling **59**, 3370–3388 (2019).

[8] L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini, "Recipe for a general, powerful, scalable graph transformer," Advances in Neural Information Processing Systems **35**, 14501–14515 (2022).

[9] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, "Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery," Chemical Reviews **119**, 10520–10594 (2019), publisher: American Chemical Society.

[10] M. M. Li, K. Huang, and M. Zitnik, "Graph representation learning in biomedicine and healthcare," Nature Biomedical Engineering **6**, 1353–1369 (2022).

[11] F. Sestak, L. Schneckenreiter, S. Hochreiter, A. Mayr, and G. Klambauer, "VN-EGNN: Equivariant graph neural networks with virtual nodes enhance protein binding site identification," in *ELLIS Machine Learning for Molecules Workshop 2023* (2023).

[12] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1: an extensible neural network potential with dft accuracy at force field computational cost," Chemical science **8**, 3192–3203 (2017).

[13] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csányi, "Mace: Higher order equivariant message passing neural networks for fast and accurate force fields," Advances in Neural Information Processing Systems **35**, 11423–11436 (2022).

[14] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," Nature Communications **13**, 2453 (2022).

[15] C. McGill, M. Forsuelo, Y. Guan, and W. H. Green, "Predicting infrared spectra with message passing neural networks," Journal of Chemical Information and Modeling **61**, 2594–2609 (2021).

[16] K. Singh, J. Munchmeyer, L. Weber, U. Leser, and A. Bande, "Graph neural networks for learning molecular excitation spectra," Journal of Chemical Theory and Computation **18**, 4408–4417 (2022).

[17] Z. Yang, M. Chakraborty, and A. D. White, "Predicting chemical shifts with graph neural networks," Chemical science **12**, 10802–10809 (2021).

[18] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen, "A graph-convolutional neural network model for the prediction of chemical reactivity," Chemical science **10**, 370–377 (2019).

[19] S. Chen and Y. Jung, "A generalized-template-based graph neural network for accurate organic reactivity prediction," Nature Machine Intelligence **4**, 772–780 (2022).

[20] B. Zhang, X. Zhang, W. Du, Z. Song, G. Zhang, G. Zhang, Y. Wang, X. Chen, J. Jiang, and Y. Luo, "Chemistry-informed molecular graph as reaction descriptor for machine-learned retrosynthesis planning," Proceedings of the National Academy of Sciences **119**, e2212711119 (2022).

[21] E. V. Konstantinova and V. A. Skorobogatov, "Molecular Hypergraphs: The New Representation of Nonclassical Molecular Structures with Polycentric Delocalized Bonds," Journal of Chemical Information and Computer Sciences **35**, 472–478 (1995), publisher: American Chemical Society.

[22] M. Skvortsova, "Molecular Graphs and Molecular Hypergraphs of Organic Compounds: Comparative Analysis," Journal of Medicinal and Chemical Sciences **4**, 452–465 (2021), publisher: Sami Publishing Company (SPC).

[23] D. W. Szczepanik, M. Andrzejak, K. Dyduch, E. Żak, M. Makowski, G. Mazur, and J. Mrozek, "A uniform approach to the description of multicenter bonding," Physical Chemistry Chemical Physics **16**, 20514–20523 (2014).

[24] F. Feixas, E. Matito, J. Poater, and M. Solà, "Understanding Conjugation and Hyperconjugation from Electronic Delocalization Measures," The Journal of Physical Chemistry A **115**, 13104–13113 (2011), publisher: American Chemical Society.

[25] G. Merino, A. Vela, and T. Heine, "Description of Electron Delocalization via the Analysis of Molecular Fields," Chemical Reviews **105**, 3812–3841 (2005), publisher: American Chemical Society.

[26] R. Liao, "Interpreting the electronic structure of the hydrogen-bridge bond in b 2 h 6 through a hypothetical reaction," Structural Chemistry **23**, 525–527 (2012).

[27] S. Bai, F. Zhang, and P. H. Torr, "Hypergraph convolution and hypergraph attention," Pattern Recognition **110**, 107637 (2021).

[28] A. Antelmi, G. Cordasco, M. Polato, V. Scarano, C. Spagnuolo, and D. Yang, "A Survey on Hypergraph Representation Learning," ACM Computing Surveys (2023), 10.1145/3605776, just Accepted.

[29] R. Aponte, R. A. Rossi, S. Guo, J. Hoffswell, N. Lipka, C. Xiao, G. Chan, E. Koh, and N. Ahmed, "A Hypergraph Neural Network Framework for Learning Hyperedge-Dependent Node Embeddings," (2022), arXiv:2212.14077 [cs].

[30] P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, and T. Laino, "Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy," Chemical science **11**, 3316–3325 (2020).

[31] K. M. Saifuddin, B. Bumgardner, F. Tanvir, and E. Akbas, "HyGNN: Drug-Drug Interaction Prediction via Hypergraph Neural Network," in *2023 IEEE 39th International Conference on Data Engineering (ICDE)* (2023) pp. 1503–1516, iSSN: 2375-026X.

[32] K. A. Murgas, E. Saucan, and R. Sandhu, "Hypergraph geometry reflects higher-order dynamics in protein interaction networks," Scientific Reports **12**, 20879 (2022).

[33] H. Kajino, "Molecular hypergraph grammar with its application to molecular optimization," in *International Conference on Machine Learning* (PMLR, 2019) pp. 3183–3191.

[34] J. Jo, J. Baek, S. J. Hwang, D. Kim, M. Kang, and S. Lee, "Edge representation learning with hypergraphs," in *Neural Information Processing Systems* (Advances in Neural Information Processing Systems, 2021).

[35] E. V. Konstantinova and V. A. Skorobogatov, "Molecular hypergraphs: the new representation of nonclassical molecular structures with polycentric delocalized bonds," Journal of chemical information and computer sciences **35**, 472–478 (1995).

[36] E. V. Konstantinova and V. A. Skorobogatov, "Application of hypergraph theory in chemistry," Discrete Mathematics **235**, 365–383 (2001).

[37] O. P. Dimitriev, "Dynamics of excitons in conjugated molecules and organic semiconductor systems," Chemical Reviews **122**, 8487–8593 (2022).

[38] H. Bronstein, C. B. Nielsen, B. C. Schroeder, and I. McCulloch, "The role of chemical design in the performance of organic semiconductors," Nature Reviews Chemistry **4**, 66–77 (2020).

[39] P. C. St John, C. Phillips, T. W. Kemper, A. N. Wilson, Y. Guan, M. F. Crowley, M. R. Nimlos, and R. E. Larsen, "Message-passing neural networks for high-throughput polymer screening," The Journal of chemical physics **150** (2019).

[40] V. Bhat, P. Sornberger, B. S. S. Pokuri, R. Duke, B. Ganapathysubramanian, and C. 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).

[41] C. Lu, Q. Liu, Q. Sun, C.-Y. Hsieh, S. Zhang, L. Shi, and C.-K. Lee, "Deep learning for optoelectronic properties of organic semiconductors," The Journal of Physical Chemistry C **124**, 7048–7060 (2020).

[42] S. Nagasawa, E. Al-Naamani, and A. Saeki, "Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest," The Journal of Physical Chemistry Letters **9**, 2639–2646 (2018).

[43] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions," (2017), arXiv:1706.08566 [physics, stat].

[44] S.-Y. Yang, "Pharmacophore modeling and applications in drug discovery: challenges and recent advances," Drug discovery today **15**, 444–450 (2010).

[45] H. E. Webel, T. B. Kimber, S. Radetzki, M. Neuenschwander, M. Nazaré, and A. Volkamer, "Revealing cytotoxic substructures in molecules using deep learning," Journal of computer-aided molecular design **34**, 731–746 (2020).

[46] P. Wang, S. Yang, Y. Liu, Z. Wang, and P. Li, "Equivariant Hypergraph Diffusion Neural Operators," (2022), arXiv:2207.06680 [cs].

[47] T. Wei, Y. You, T. Chen, Y. Shen, J. He, and Z. Wang, "Augmentations in Hypergraph Contrastive Learning: Fabricated and Generative," (2022), arXiv:2210.03801 [cs].

[48] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open graph benchmark: datasets for machine learning on graphs," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20 (Curran Associates Inc., Red Hook, NY, USA, 2020) pp. 22118–22133.

[49] W. Hu, M. Fey, H. Ren, M. Nakata, Y. Dong, and J. Leskovec, "OGB-LSC: A Large-Scale Challenge for Machine Learning on Graphs," (2021), arXiv:2103.09430 [cs].

[50] M. Nakata and T. Shimazaki, "PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry," Journal of Chemical Information and Modeling **57**, 1300–1308 (2017), publisher: American Chemical Society.

[51] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations* (2017).

[52] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" (2019), arXiv:1810.00826 [cs, stat].

[53] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," (2018), arXiv:1710.10903 [cs, stat].

[54] S. Brody, U. Alon, and E. Yahav, "How Attentive are Graph Attention Networks?" (2022), arXiv:2105.14491 [cs].

[55] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural Message Passing for Quantum Chemistry," (2017), arXiv:1704.01212 [cs].

[56] J. Kim, D. Nguyen, S. Min, S. Cho, M. Lee, H. Lee, and S. Hong, "Pure transformers are powerful graph learners," Advances in Neural Information Processing Systems **35**, 14582–14595 (2022).

[57] Z. ZHANG, Q. Liu, S. Zhang, C.-Y. Hsieh, L. Shi, and C.-K. Lee, "Graph self-supervised learning for optoelectronic properties of organic semiconductors," in *ICML 2022 2nd AI for Science Workshop* (2022).

[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Inc., 2019).

[59] M. Fey and J. E. Lenssen, "Fast Graph Representation Learning with PyTorch Geometric," (2019), arXiv:1903.02428 [cs, stat].