

RESEARCH ARTICLE | APRIL 10 2024

Molecular hypergraph neural networks ^{EP}

Special Collection: [2024 JCP Emerging Investigators Special Collection](#)

Junwu Chen ; Philippe Schwaller  



J. Chem. Phys. 160, 144307 (2024)

<https://doi.org/10.1063/5.0193557>

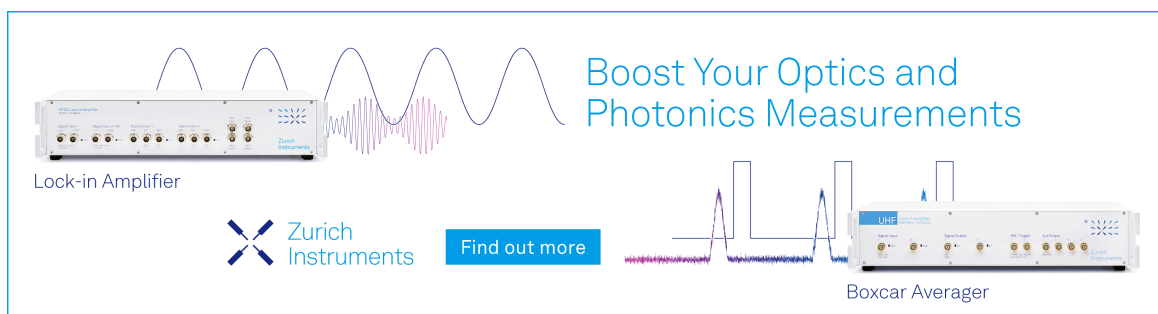


View
Online




Export
Citation

01 May 2024 16:20:34



Boost Your Optics and
Photonics Measurements

Lock-in Amplifier

 Zurich
Instruments

[Find out more](#)

Boxcar Averager

Molecular hypergraph neural networks

Cite as: J. Chem. Phys. **160**, 144307 (2024); doi: [10.1063/5.0193557](https://doi.org/10.1063/5.0193557)

Submitted: 22 December 2023 • Accepted: 14 March 2024 •

Published Online: 10 April 2024



View Online



Export Citation



CrossMark

Junwu Chen^{1,2,a)}  and Philippe Schwaller^{1,2,b)} 

AFFILIATIONS

¹Laboratory of Artificial Chemical Intelligence (LIAC), Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

²National Centre of Competence in Research (NCCR) Catalysis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Note: This paper is part of the 2024 JCP Emerging Investigators Special Collection.

^{a)}Electronic mail: junwu.chen@epfl.ch

^{b)}Author to whom correspondence should be addressed: philippe.schwaller@epfl.ch

ABSTRACT

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, such as multi-center bonds and conjugated structures. To tackle this challenge, we introduce molecular hypergraphs and propose Molecular Hypergraph Neural Networks (MHNNs) 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 organic photovoltaic, OCELOT chromophore v1, 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.

© 2024 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0193557>

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 of effectively representing higher order connections.^{11,21,22} A substantial number of molecules have delocalized bonds, such as multi-center bonds²³ 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 Fig. 1(a), two B atoms and one H atom share two electrons to form a three-center-two-electron bond, which cannot be represented by a pairwise edge.²⁶ Similarly, conjugated organic molecules like porphyrin in Fig. 1(b) 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.

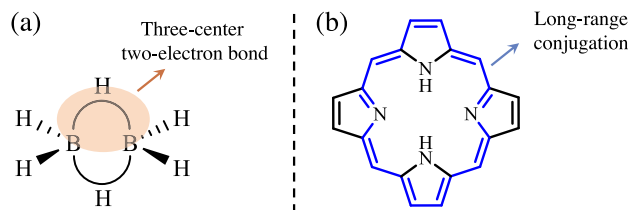


FIG. 1. (a) Dimeric borane structure and its three-center-two-electron bond (B–H–B). (b) Porphyrin structure and its long-range conjugated bond.

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 (HGNNs) 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 HGNNs 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 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 the 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 $e = \{v_1, \dots, v_{|e|}\}$ 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 substantially 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., HOMO–LUMO gap).³⁸ Moreover, hyperedges could be defined by pharmacophores⁴⁴ or toxicophores⁴⁵ for predicting molecular activity or toxicity, as pharmacophores/toxicophores are crucial components within molecules determining activity/toxicity. In this work, we show an example of using molecular hypergraphs to describe conjugated molecules [Fig. 2(a)], 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 12 nodes, 12 two-order hyperedges, and one six-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 (MHNNs) based on bipartite representations of hypergraphs, which can efficiently operate on hypergraphs with hyperedges of various orders [Figs. 2(b) and 2(c)].

The molecular hypergraph is initially transformed into an equivalent bipartite graph [Fig. 2(b)], wherein two distinct sets of vertices denote the nodes and hyperedges of the molecular hypergraph, respectively. The message passing of MHNN 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 ($1 \leq t \leq T$) step message passing, the hidden states $l_e^{(t)}$ of each hyperedge are updated based on the messages $m_{v \rightarrow e}^{(t)}$ from the connected nodes ($v \in e$) according to

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

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

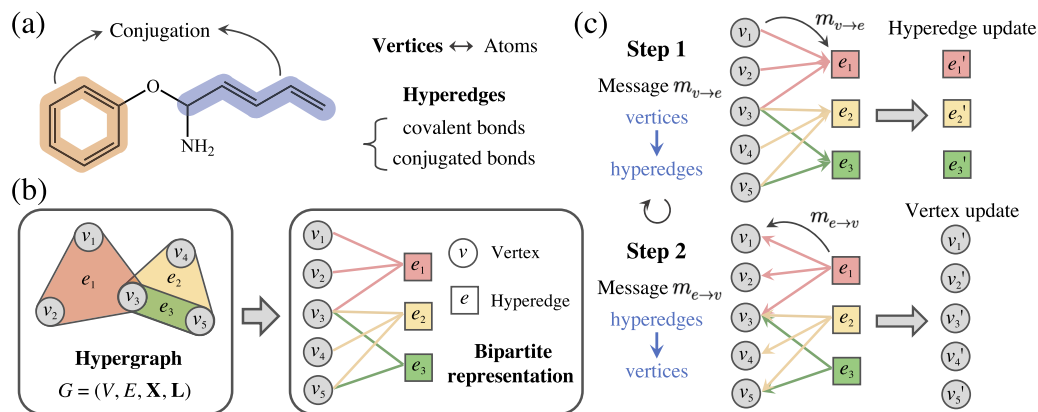


FIG. 2. (a) The method of constructing molecular hypergraphs for conjugated molecules. (b) The conversion from a hypergraph to an equivalent bipartite graph. (c) The message-passing method of our MHNN model.

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

$$m_{e \rightarrow v}^{(t)} = \sum_{e: v \in e} f_3(I_e^{(t)}, h_v^{(t-1)}), \quad (3)$$

$$h_v^{(t)} = f_4(h_v^{(t-1)}, m_{e \rightarrow v}^{(t)}), \quad (4)$$

where $h_v^{(0)}$ and $I_e^{(0)}$ are derived from initial atom features and bond features (Appendix B), respectively. 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

$$\hat{y} = \text{MLP}\left(\sum_{v \in G} h_v^{(T)}, \sum_{e \in G} I_e^{(T)}\right), \quad (5)$$

where $\text{MLP}(\cdot)$ is a multi-layer perceptron. The output \hat{y} is the prediction target of MHNN, which can be a scalar or a vector, depending on whether the prediction task is for a single property or multiple properties. In this work, four MLPs are used to act as update functions (f_1 , f_2 , f_3 , and f_4). The schematic diagram of MHNN architecture is shown in Fig. 3 and Algorithm 1.

C. Input features

For 2D GNN baselines, the atoms features and bond features designed by Open Graph Benchmark (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 3D GNN baselines, only atomic numbers are used as the initial node feature. The RDKit⁴⁹ software was used to calculate atomic features and bond features from simplified molecular-input line-entry system (SMILES) strings and served as input to MHNN and baseline models. The results of all models come from single-target regression prediction tasks. More details are listed in Appendix B.

D. Datasets

The OPV dataset,³⁹ named the organic photovoltaic dataset, contains 90 823 unique molecules (monomers and soluble small molecules) and their SMILES strings, 3D geometries, and optoelectronic properties from density functional theory (DFT) calculations. OPV has four molecular tasks for monomers, the energy of the highest occupied molecular orbital (ϵ_{HOMO}), the lowest unoccupied molecular orbital (ϵ_{LUMO}), the HOMO–LUMO gap ($\Delta\epsilon$), and the spectral overlap I_{overlap} . In addition, OPV has four polymeric tasks, the polymer ϵ_{HOMO} , polymer ϵ_{LUMO} , polymer gap $\Delta\epsilon$, and optical LUMO O_{LUMO} .³⁹ The more detailed descriptions and calculation methods of the above properties can be found in Ref. 39.

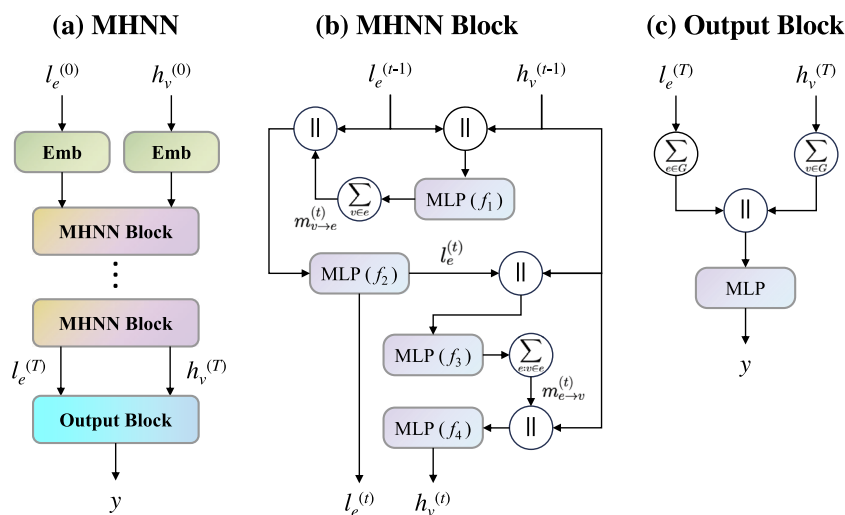
The OCELOT chromophore v1 (OCELOTv1) dataset⁴⁰ comprises about 25 000 organic π -conjugated molecules, along with their optoelectronic and reaction characteristics calculated by precise DFT or time-dependent DFT (TD-DFT) methods. The dataset encompasses 15 molecular properties: vertical (VIE) and adiabatic (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.

PCQM4Mv2⁵⁰ is based on the PubChemQC project⁵¹ 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 OCELOTv1 dataset. The experimental results are derived from five separate runs using different random seeds, except for PCQM4Mv2, which is based on one single random seed run (Table I).

III. RESULTS AND DISCUSSION

In this section, we initially assessed the predictive performance of MHNN on optoelectronic properties across three datasets.

ALGORITHM 1. Algorithm of MHNN.**Require:** Molecular hypergraph $G = (V, E, \mathbf{H}, \mathbf{L})$ 1: 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.2: **for** $t = 1, 2, \dots, T$ **do**3: Send messages from V to E for all $e \in E$: $m_{v \rightarrow e}^{(t)} = \sum_{v \in e} f_1([h_v^{(t-1)}, l_e^{(t-1)}])$ 4: Update hyperedge embeddings $l_e^{(t)} = f_2([l_e^{(t-1)}, m_{v \rightarrow e}^{(t)}])$ 5: Send messages from E to V : $m_{e \rightarrow v}^{(t)} = \sum_{e: v \in e} f_3([l_e^{(t)}, h_v^{(t-1)}])$ 6: Update node embeddings $h_v^{(t)} = f_4([h_v^{(t-1)}, m_{e \rightarrow v}^{(t)}])$ 7: **end for**8: Hypergraph embedding from nodes: $g_v = \sum_{v \in G} h_v^{(T)}$ 9: Hypergraph embedding from hyperedges: $g_e = \sum_{e \in G} l_e^{(T)}$, $|e| > 2$ 10: $\hat{y} = \text{MLP}([g_v, g_e])$ **Ensure:** \hat{y} **FIG. 3.** The MHNN architecture. \parallel denotes concatenation. The embeddings of nodes and hyperedges are updated in multiple MHNN blocks that can share parameters or not. The final embeddings of nodes and hyperedges are passed into an output block to generate predictions.**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

Among them, the OPV³⁹ and OCELOTv1⁴⁰ 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 on conjugated structure and its prediction performance for optoelectronic properties. As shown in Fig. 4(a), 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 500–1000, whereas the OCELOTv1 dataset shows a concentration in the range of 200–400. 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. 4(b), molecules with larger conjugated structures are present in the OPV dataset compared to the OCELOTv1 dataset. The num-

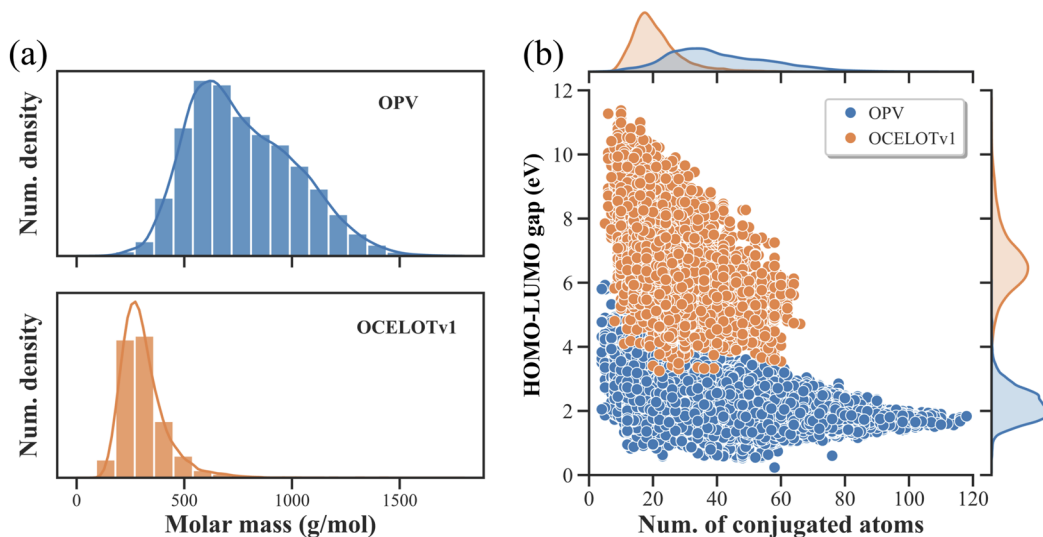


FIG. 4. (a) Distribution of molecular weights for OPV and OCELOTv1 datasets. (b) Distribution of HOMO-LUMO gap and atomic number of conjugated structures for OPV and OCELOTv1 datasets.

ber 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 HOMO-LUMO gaps (~ 1.9 eV) compared to the OCELOTv1 dataset (~ 6.2 eV). The distribution without obvious regularity also demonstrates the complex relationship between the photoelectric properties and conjugated structures.

B. Performance on OPV dataset

For the OPV dataset, we compared MHNN with multiple baselines: Graph Convolutional Network (GCN),⁵² Graph Isomorphism Networks (GIN),⁵³ Graph Attention Network (GAT),⁵⁴ GATv2,⁵⁵ MPNN,⁵⁶ 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 SchNet⁴³ that 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 important for these properties.³⁹ However, MHNN outperforms all baselines on three tasks ($\Delta\epsilon$, ϵ_{HOMO} , and ϵ_{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 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, SchNet⁴³ 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 seven out of eight tasks compared to baselines, which demonstrates the significance of molecular hypergraphs and the excellent performance of MHNN for property prediction of conjugated molecules. The prominence of MHNN could come from the molecular hypergraph and its algorithm. The molecular hypergraph can help MHNN understand that conjugated structures are important components, and guide MHNN to learn the correlation between conjugated structures and photoelectric properties. Moreover, MHNN's message passing occurs between nodes and hyperedges, which involves atoms, chemical bonds, and conjugated structures. Then, the embeddings of both nodes and hyperedges representing conjugate structures are used to compute the model output. Therefore, MHNN's algorithm essentially enables it to learn the relationship between conjugated structures and optoelectronic properties, thereby enhancing its predictive performance.

C. Performance on OCELOTv1 dataset

All models from the original paper⁴⁰ 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 an FFN.⁴⁰ More details about the baseline models can be found in Ref. 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, and FFN)

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 Ref. 39. Standard deviation is used to evaluate the error in test results. The best results for each target are marked in bold.

Methods	Molecular				Polymer			
	$\Delta\epsilon$	ϵ_{HOMO}	ϵ_{LUMO}	$I_{overlap}$	$\Delta\epsilon$	ϵ_{HOMO}	ϵ_{LUMO}	O_{LUMO}
GCN	66.7 ± 1.0	38.7 ± 0.8	53.8 ± 1.8	268.1 ± 3.9	77.0 ± 0.4	54.1 ± 0.6	62.2 ± 0.8	61.4 ± 0.7
GIN	48.8 ± 0.6	29.3 ± 0.2	38.9 ± 0.3	187.3 ± 2.6	67.0 ± 1.2	48.4 ± 0.3	54.5 ± 0.7	53.7 ± 1.1
GAT	54.4 ± 1.1	33.7 ± 0.7	43.1 ± 1.3	203.8 ± 6.9	72.1 ± 1.4	51.7 ± 1.2	58.8 ± 0.7	57.8 ± 0.9
GATv2	57.2 ± 2.1	32.8 ± 1.4	44.2 ± 1.8	199.8 ± 1.2	72.8 ± 0.9	51.8 ± 0.5	57.5 ± 0.8	58.1 ± 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	28.5 ± 0.2	21.9 ± 0.1	21.1 ± 0.3	112.9 ± 0.8	56.9 ± 0.3	46.0 ± 0.6	45.3 ± 0.5	44.7 ± 0.2

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 Ref. 40. Standard deviation is used to evaluate the error in test results. The best results for each target are marked in bold.

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	0.309 ± 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	0.259 ± 0.002
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	0.522 ± 0.009
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	0.176 ± 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	0.163 ± 0.003
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 ± 0.001
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	0.100 ± 0.002
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	0.139 ± 0.002
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	0.126 ± 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	0.093 ± 0.002
SOS1	0.307 ± 0.006	0.275 ± 0.004	0.307 ± 0.002	0.282 ± 0.003	0.252 ± 0.017	0.249 ± 0.013	0.243 ± 0.003
SOT1	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 ± 0.003

TABLE IV. Validate MAE results of MHNN and other message-passing GNN baselines on the PCQM4Mv2. The results of baselines are from Refs. 50 and 56. This dataset does not publish its test set. VN represents the use of virtual nodes to improve performance. The best result is marked in bold.

Model	Parameters (M)	Validate MAE (eV)
GCN	2.0	0.1379
GIN	3.8	0.1195
GAT	6.7	0.1302
GCN-VN	4.9	0.1153
GAT-VN	6.7	0.1192
MHNN	2.1	0.1120

using molecular descriptors. However, the models using molecular descriptors show superior performance than MPNN in the tasks, such as HOMO, H-L, and HR. Moreover, with the assistance of extra molecular descriptors, the MPNN + MolDes model demon-

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

D. Performance on PCQM4Mv2 dataset

To explore the learning ability on a 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

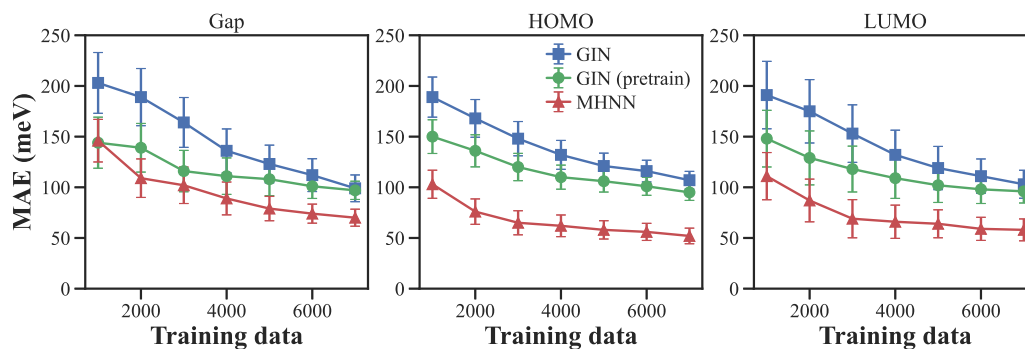


FIG. 5. The test results of different models on the HOMO–LUMO gap, HOMO and LUMO tasks of OPV dataset under different amounts of training data. The green lines represent the results of pretrained GIN by self-supervised learning,⁵⁷ while the blue and red lines show the results from GIN and MHNN without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

is the HOMO–LUMO gap, which is one of the optoelectronic properties. As given 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 the OPV dataset under the same data partition. All 80 823 unlabeled molecules in the training set were used to pretrain the GIN model using the 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 Fig. 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 ϵ_{HOMO} and ϵ_{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, and PCQM4Mv2), MHNN exhibits superior performance to the baselines on most tasks. Impressively, even in the absence of 3D geometric information, MHNN surpasses SchNet that relies on atom positions. Moreover, MHNN demonstrates higher data efficiency

compared to pretrained models, making it valuable for applications where labeled data are scarce. When specific parts of molecular structures substantially contribute to the target properties, MHNN can use hyperedges to describe and learn these higher order interactions. For instance, pharmacophores/toxicophores are crucial components within molecules that determine activity/toxicity, and thus, they can be represented by hyperedges to facilitate the activity/toxicity prediction of MHNN. Molecular hypergraphs and MHNN provide a new strategy for property prediction involving higher order connections.

ACKNOWLEDGMENTS

This article was created as part of NCCR Catalysis (Grant No. 180544), a National Center of Competence in Research funded by the Swiss National Science Foundation.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Junwu Chen: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Resources (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal). **Philippe Schwaller:** Conceptualization (equal); Funding acquisition (equal); Methodology (equal); Project administration (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The code of MHNN and baseline models for optoelectronic property prediction on the OPV,³⁹ OCELOTv1,⁴⁰

and PCQM4Mv2⁵⁰ datasets can be found on GitHub at <https://github.com/schwallergroup/mhnn>.

This work uses three open-source datasets, OPV,³⁹ OCELOTv1,⁴⁰ and PCQM4Mv2.⁵⁰ The data that support the findings of this study are openly available on GitHub at <https://github.com/schwallergroup/mhnn>, including the scripts to download and process the datasets.

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 central processing unit (CPU) and graphics processing unit (GPU) architectures. This included a combination of NVidia V100 (32 GB) and RTX3090 (24 GB) 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 Refs. 39 and 50.

APPENDIX B: INPUT FEATURES

Tables V–VII describe the input features for atoms, pair-wise edges, and hyperedges.

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

Feature	Description
Atom type	Type of atom (e.g., 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, sp ² , sp ³ , sp ³ d, or sp ³ d ²
Aromaticity	Whether this atom is part of an aromatic system
Is in ring	Whether the atom is in a ring

TABLE VI. 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 VII. Using bond type as the hyperedge feature of MHNN.

Edge order	Feature
= 2	Bond type: Single, double, triple, or aromatic
> 2	Conjugated bonds

REFERENCES

- D. K. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, “Convolutional networks on graphs for learning molecular fingerprints,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems* (MIT Press, 2015), Vol. 2, pp. 2224–2232.
- J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in *Proceedings of the 34th International Conference on Machine Learning* (PMLR, 2017), Vol. 70, pp. 1263–1272, see <http://proceedings.mlr.press/v70/gilmer17a.html>.
- J. Gasteiger, J. Groß, and S. Günnemann, “Directional message passing for molecular graphs,” *arXiv:2003.03123* (2020).
- P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoessel, H. Schopmans, T. Sommer, and P. Friederich, “Graph neural networks for materials science and chemistry,” *Commun. Mater.* **3**, 93 (2022).
- K. Atz, F. Grisoni, and G. Schneider, “Geometric deep learning on molecular representations,” *Nat. Mach. Intell.* **3**, 1023–1032 (2021).
- 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,” *Nat. Mach. Intell.* **4**, 127–134 (2022).
- 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,” *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
- L. Rampásek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini, “Recipe for a general, powerful, scalable graph transformer,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems* (MIT Press, 2022), Vol. 35, pp. 14501–14515.
- X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, “Concepts of artificial intelligence for computer-assisted drug discovery,” *Chem. Rev.* **119**, 10520–10594 (2019).
- M. M. Li, K. Huang, and M. Zitnik, “Graph representation learning in biomedicine and healthcare,” *Nat. Biomed. Eng.* **6**, 1353–1369 (2022).
- F. Sestak, L. Schneckenreiter, S. Hochreiter, A. Mayr, and G. Klambauer, “VN-EINN: Equivariant graph neural networks with virtual nodes enhance protein binding site identification,” in *ELLIS Machine Learning for Molecules Workshop*, 2023.
- J. S. Smith, O. Isayev, and A. E. Roitberg, “ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost,” *Chem. Sci.* **8**, 3192–3203 (2017).
- 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,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems* (MIT Press, 2022), Vol. 35, pp. 11423–11436.
- 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,” *Nat. Commun.* **13**, 2453 (2022).
- C. McGill, M. Forsuelo, Y. Guan, and W. H. Green, “Predicting infrared spectra with message passing neural networks,” *J. Chem. Inf. Model.* **61**, 2594–2609 (2021).
- K. Singh, J. Munchmeyer, L. Weber, U. Leser, and A. Bande, “Graph neural networks for learning molecular excitation spectra,” *J. Chem. Theory Comput.* **18**, 4408–4417 (2022).
- Z. Yang, M. Chakraborty, and A. D. White, “Predicting chemical shifts with graph neural networks,” *Chem. Sci.* **12**, 10802–10809 (2021).
- 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,” *Chem. Sci.* **10**, 370–377 (2019).
- S. Chen and Y. Jung, “A generalized-template-based graph neural network for accurate organic reactivity prediction,” *Nat. Mach. Intell.* **4**, 772–780 (2022).
- 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,” *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2212711119 (2022).

- ²¹E. V. Konstantinova and V. A. Skorobogatov, "Molecular hypergraphs: The new representation of nonclassical molecular structures with polycyclic delocalized bonds," *J. Chem. Inf. Comput. Sci.* **35**, 472–478 (1995).
- ²²M. Skvortsova, "Molecular graphs and molecular hypergraphs of organic compounds: Comparative analysis," *J. Med. Chem. Sci.* **4**, 452–465 (2021).
- ²³D. W. Szczepanik, M. Andrzejak, K. Dyduch, E. Żak, M. Makowski, G. Mazur, and J. Mrozek, "A uniform approach to the description of multicenter bonding," *Phys. Chem. Chem. Phys.* **16**, 20514–20523 (2014).
- ²⁴F. Feixas, E. Matito, J. Poater, and M. Solà, "Understanding conjugation and hyperconjugation from electronic delocalization measures," *J. Phys. Chem. A* **115**, 13104–13113 (2011).
- ²⁵G. Merino, A. Vela, and T. Heine, "Description of electron delocalization via the analysis of molecular fields," *Chem. Rev.* **105**, 3812–3841 (2005).
- ²⁶R. Liao, "Interpreting the electronic structure of the hydrogen-bridge bond in B₂H₆ through a hypothetical reaction," *Struct. Chem.* **23**, 525–527 (2012).
- ²⁷S. Bai, F. Zhang, and P. H. Torr, "Hypergraph convolution and hypergraph attention," *Pattern Recognit.* **110**, 107637 (2021).
- ²⁸A. Antelmi, G. Cordasco, M. Polato, V. Scarano, C. Spagnuolo, and D. Yang, "A survey on hypergraph representation learning," *ACM Comput. Surv.* **56**, 1 (2023).
- ²⁹R. 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," [arXiv:2212.14077](https://arxiv.org/abs/2212.14077) [cs] (2022).
- ³⁰P. 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," *Chem. Sci.* **11**, 3316–3325 (2020).
- ³¹K. 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)* (IEEE, 2023), pp. 1503–1516.
- ³²K. A. Murgas, E. Saucan, and R. Sandhu, "Hypergraph geometry reflects higher-order dynamics in protein interaction networks," *Sci. Rep.* **12**, 20879 (2022).
- ³³H. Kajino, "Molecular hypergraph grammar with its application to molecular optimization," in *Proceedings of the 36th International Conference on Machine Learning (PMLR, 2019)*, Vol. 97, pp. 3183–3191, see <http://proceedings.mlr.press/v97/kajino19a.html>.
- ³⁴J. Jo, J. Baek, S. J. Hwang, D. Kim, M. Kang, and S. Lee, "Edge representation learning with hypergraphs," in *Advances in Neural Information Processing Systems*, 2021.
- ³⁵E. V. Konstantinova and V. A. Skorobogatov, "Molecular structures of organoelement compounds and their representation as labeled molecular hypergraphs," *J. Struct. Chem.* **39**(2), 268–276 (1998).
- ³⁶E. V. Konstantinova and V. A. Skorobogatov, "Application of hypergraph theory in chemistry," *Discrete Math.* **235**, 365–383 (2001).
- ³⁷O. P. Dimitriev, "Dynamics of excitons in conjugated molecules and organic semiconductor systems," *Chem. Rev.* **122**, 8487–8593 (2022).
- ³⁸H. Bronstein, C. B. Nielsen, B. C. Schroeder, and I. McCulloch, "The role of chemical design in the performance of organic semiconductors," *Nat. Rev. Chem.* **4**, 66–77 (2020).
- ³⁹P. 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," *J. Chem. Phys.* **150**, 234111 (2019).
- ⁴⁰V. 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," *Chem. Sci.* **14**, 203–213 (2023).
- ⁴¹C. Lu, Q. Liu, Q. Sun, C.-Y. Hsieh, S. Zhang, L. Shi, and C.-K. Lee, "Deep learning for optoelectronic properties of organic semiconductors," *J. Phys. Chem. C* **124**, 7048–7060 (2020).
- ⁴²S. Nagasawa, E. Al-Naamani, and A. Saeki, "Computer-aided screening of conjugated polymers for organic solar cell: Classification by random forest," *J. Phys. Chem. Lett.* **9**, 2639–2646 (2018).
- ⁴³K. 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," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (MIT Press, 2017), Vol. 30, pp. 992–1002.
- ⁴⁴S.-Y. Yang, "Pharmacophore modeling and applications in drug discovery: Challenges and recent advances," *Drug Discovery Today* **15**, 444–450 (2010).
- ⁴⁵H. E. Weibel, T. B. Kimber, S. Radetzki, M. Neuenschwander, M. Nazaré, and A. Volkamer, "Revealing cytotoxic substructures in molecules using deep learning," *J. Comput.-Aided Mol. Des.* **34**, 731–746 (2020).
- ⁴⁶P. Wang, S. Yang, Y. Liu, Z. Wang, and P. Li, "Equivariant hypergraph diffusion neural operators," [arXiv:2207.06680](https://arxiv.org/abs/2207.06680) [cs] (2022).
- ⁴⁷T. Wei, Y. You, T. Chen, Y. Shen, J. He, and Z. Wang, "Augmentations in hypergraph contrastive learning: Fabricated and generative," [arXiv:2210.03801](https://arxiv.org/abs/2210.03801) [cs] (2022).
- ⁴⁸W. 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, 2020), pp. 22118–22133.
- ⁴⁹G. Landrum, P. Tosco, B. Kelley, Ric. Sriniker, Gedeck, R. Vianello, D. Cosgrove, N. Schneider, E. Kawashima, D. N. A. Dalke, G. Jones, B. Cole, M. Swain, S. Turk, A. Savelyev, A. Vaucher, M. Wójcikowski, I. Take, D. Probst, K. Ujihara, V. F. Scalfani, G. Godin, A. Pahl, F. Berenger, J. L. Varjo, J. P. Strets, and D. Gavid, "RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling," Version: release_2022_09_3, 2022.
- ⁵⁰W. Hu, M. Fey, H. Ren, M. Nakata, Y. Dong, and J. Leskovec, "OGB-LSC: A large-scale challenge for machine learning on graphs," [arXiv:2103.09430](https://arxiv.org/abs/2103.09430) [cs] (2021).
- ⁵¹M. Nakata and T. Shimazaki, "PubChemQC project: A large-scale first-principles electronic structure database for data-driven chemistry," *J. Chem. Inf. Model.* **57**, 1300–1308 (2017).
- ⁵²T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.
- ⁵³K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," [arXiv:1810.00826](https://arxiv.org/abs/1810.00826) [cs, stat] (2019).
- ⁵⁴P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," [arXiv:1710.10903](https://arxiv.org/abs/1710.10903) [cs, stat] (2018).
- ⁵⁵S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?," [arXiv:2105.14491](https://arxiv.org/abs/2105.14491) [cs] (2022).
- ⁵⁶J. Kim, D. Nguyen, S. Min, S. Cho, M. Lee, H. Lee, and S. Hong, "Pure transformers are powerful graph learners," in *Proceedings of the 36th International Conference on Neural Information Processing Systems* (MIT Press, 2022), Vol. 35, pp. 14582–14595.
- ⁵⁷Z. 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.
- ⁵⁸A. 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* (Curran Associates, Inc., 2019), Vol. 32.
- ⁵⁹M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch geometric," [arXiv:1903.02428](https://arxiv.org/abs/1903.02428) [cs, stat] (2019).