

Bi-level Contrastive Learning for Knowledge-Enhanced Molecule Representations

Pengcheng Jiang¹, Cao Xiao², Tianfan Fu³, Parminder Bhatia², Taha Kass-Hout²,
Jimeng Sun¹, Jiawei Han¹

¹University of Illinois Urbana Champaign

²GE HealthCare

³Rensselaer Polytechnic Institute

Abstract

Molecular representation learning is vital for various downstream applications, including the analysis and prediction of molecular properties and side effects. While Graph Neural Networks (GNNs) have been a popular framework for modeling molecular data, they often struggle to capture the full complexity of molecular representations. In this paper, we introduce a novel method called GODE, which accounts for the dual-level structure inherent in molecules. Molecules possess an intrinsic graph structure and simultaneously function as nodes within a broader molecular knowledge graph. GODE integrates individual molecular graph representations with multi-domain biochemical data from knowledge graphs. By pre-training two GNNs on different graph structures and employing contrastive learning, GODE effectively fuses molecular structures with their corresponding knowledge graph substructures. This fusion yields a more robust and informative representation, enhancing molecular property predictions by leveraging both chemical and biological information. When fine-tuned across 11 chemical property tasks, our model significantly outperforms existing benchmarks, achieving an average ROC-AUC improvement of 12.7% for classification tasks and an average RMSE/MAE improvement of 34.4% for regression tasks. Notably, GODE surpasses the current leading model in property prediction, with advancements of 2.2% in classification and 7.2% in regression tasks.

Introduction

In recent years, there has been a significant focus on tailoring machine learning models specifically for chemical and biological data (Wang et al. 2021a; Li, Huang, and Zitnik 2022; Somnath, Bunne, and Krause 2021; Wang et al. 2023). A key challenge in this field is developing effective representations of molecular structures, which are critical for achieving accurate predictions in subsequent tasks (Yang et al. 2019; Haghghatlari et al. 2020). To address this challenge, graph neural networks (GNNs) have become a widely adopted tool for facilitating representation learning (Li et al. 2021; Hu et al. 2019). However, the conventional approach of using molecular graphs as input for GNNs may inadvertently constrain their potential for generating robust and comprehensive representations.

Molecular data, including chemical and biological datasets, exhibits a wide range of representational complexities (Tong et al. 2017; Argelaguet et al. 2020). On an individual level, molecules can naturally be represented as graphs, with atoms as nodes and bonds as edges. For collections of molecules, their interrelationships can be captured through knowledge graphs (KGs), where each molecule is represented as a unique node. Notable examples of such KGs include UMLS (Bodenreider 2004), PrimeKG (Chandak, Huang, and Zitnik 2023a), and PubChemRDF (Fu et al. 2015). Building on this observation, we hypothesize that by integrating the molecular graphs of individual molecules and the broader sub-graphs from KGs centered on these molecules, we can create a more enriched representation that could lead to more accurate and robust predictions.

Previous efforts have attempted to unify molecular structures with KGs for property prediction. For example, Ye et al. (2021) combined molecule embeddings with static KG embeddings (Bordes et al. 2013). However, these integrations often fall short of capturing the local molecular information within the KG, leading to only marginal improvements in prediction accuracy. In contrast, Fang et al. (2023) demonstrated the advantages of enhancing molecular representations through contrastive learning, supported by a specialized chemical element KG. This approach has shown more significant performance gains, underscoring the value of integrating KGs with molecular data. Our work seeks to explore novel methods for embedding biochemical knowledge graphs into molecular prediction models.

In this study, we introduce “**Graph as a Node**” (GODE), a new approach specifically designed to pre-train GNNs for molecule predictions. Our approach incorporates bi-level self-supervised tasks that target both molecular structures and their corresponding sub-graphs within the knowledge graph. By combining this strategy with contrastive learning, GODE produces more robust embeddings, leading to improved predictions of molecular properties.

Our major contributions can be summarized as follows:

- **A new paradigm for molecule knowledge integration.** Our GODE method introduces a new approach to integrating molecular structures with their corresponding KGs. This method not only produces richer and more accurate molecular representations in our specific application but also has the potential to be extended to other domains.

- **More robust molecular embeddings.** Achieving robust molecular representations is crucial for accurate and consistent predictions across diverse datasets. Our approach integrates information from multiple domains for the same molecule, leveraging shared knowledge across modalities to create more comprehensive representations. By utilizing bi-level self-supervised pre-training combined with contrastive learning, we significantly enhance the robustness and reliability of the embeddings, resulting in more precise molecular property predictions and a solid foundation for various applications.
- **Introducing a new molecular knowledge graph.** We have developed MolKG, a comprehensive knowledge graph specifically tailored to molecular data. MolKG encapsulates extensive molecular information, enabling more advanced and knowledge-driven molecular analyses.

To evaluate GODE’s performance, we conducted experiments across 11 chemical property prediction tasks. We benchmarked GODE against state-of-the-art methods, including GROVER (Rong et al. 2020), MolCLR (Wang et al. 2021a), and KANO (Fang et al. 2023). Our results demonstrate that GODE consistently outperforms these baselines, achieving improvements of 12.7% in classification tasks and 34.4% in regression tasks for molecular property prediction.

Related Works

Graph-based Molecular Representation Learning. Over the years, various streams of molecular representation methods have been proposed, including traditional fingerprint-based approaches (Rogers and Hahn 2010), SMILES string methods (Xu et al. 2024), and modern GNN methods (Jin et al. 2017, 2018; Zheng et al. 2019). While Mol2Vec (Jaeger, Fulle, and Turk 2018) adopts a molecule interpretation akin to Word2Vec for sentences (Mikolov et al. 2013), it overlooks substructure roles in chemistry. In contrast, GNN-based techniques can overcome this limitation by capturing more insightful details from aggregated sub-graphs. This advantage yields enhanced representations for chemical nodes, bonds, and entire molecules (Rong et al. 2020; Hu et al. 2019; Wang et al. 2021a). Consequently, our study adopts GNN as the foundational framework for representing molecules.

Biomedical Knowledge Graphs. Various biomedical/biochemical KGs were developed to capture interconnections among diverse entities such as genes, proteins, diseases, and drugs (Fu et al. 2015; Bodenreider 2004). Notably, PubChemRDF (Fu et al. 2015) spotlights biochemical domains, furnishing machine-readable chemical insights encompassing structures, properties, activities, and bioassays. Its subdivisions (e.g., *Compound*, *Cooccurrence*, *Descriptor*, *Pathway*) amass comprehensive chemical information. PrimeKG (Chandak, Huang, and Zitnik 2023b) is another KG that provides a multimodal view of precision medicine. Our study has a complementary focus and constructs a molecule-centric KG from those base KGs for supporting molecule property prediction tasks.

Molecular Property Predictions. We focus on molecular property prediction, an essential downstream task for chem-

ical representation learning frameworks. Three main aspects of the molecular property attract researchers: quantum mechanics properties (Yang et al. 2019; Liao et al. 2019; Gilmer et al. 2017), physicochemical properties (Shang et al. 2018; Wang et al. 2019; Bécigneul et al. 2020), and toxicity (Xu, Pei, and Lai 2017; Yuan and Ji 2020). Most of the recent works on molecular predictions are based on GNN (Duvinaud et al. 2015; Mansimov et al. 2019). However, the methods mentioned only focus on chemical structures and do not consider inter-relations among chemicals and knowledge graphs, which could improve property prediction.

Contrastive Learning in Molecular Representation. The rise of cross-modality contrastive learning (Radford et al. 2021) has increasingly influenced molecular representation approaches. Pioneering studies, such as (Stärk et al. 2022), have successfully employed contrastive learning to merge 3D and 2D molecular representations. This technique has been applied across various domains, including chemical reactions (Lee et al. 2021; Seidl et al. 2022), natural language processing (Su et al. 2022; Seidl et al. 2023), microscopy imaging (Sanchez-Fernandez et al. 2022), and chemical element knowledge (Fang et al. 2023). Distinctively, our work harnesses contrastive learning to enable knowledge transfer between KGs and molecular structures.

Fusing Knowledge Graph and Molecules. Previously, Ye et al. (2021) introduced a method that combines static KG embeddings of drugs with their structural representations for downstream tasks. However, this approach overlooks the contextual information surrounding molecular nodes, leading to only modest performance improvements. Alternatively, Wang et al. (2022) proposed a Graph-of-Graphs technique that enriches molecular graph representations. While this method enhances the graph information, it does not explore pre-training or contrastive learning strategies to align the same entity across different graph modalities. On the other hand, Fang et al. (2023) pioneered a contrastive learning-based approach that augments molecular structures with element-wise knowledge, creating an innovative graph structure that has significantly improved molecular property predictions. Unlike existing methods, GODE extracts a molecule’s sub-graph from our molecule-centric KG, offering a novel representation that effectively links molecular data with knowledge graphs.

Method

In this section, we present GODE framework. First, we define a few key concepts below.

Definition 1 (Molecule Graph) A molecule graph (MG) is a structured representation of a molecule, where atoms (or nodes) are connected by bonds (or edges). An MG G_m can be viewed as a graph structure with a set of nodes \mathcal{V}_m representing atoms and a set of edges \mathcal{E}_m representing bonds such that $G_m = (\mathcal{V}_m, \mathcal{E}_m)$.

Definition 2 (Knowledge Graph) A knowledge graph (KG) is a structured representation of knowledge in which entities (or nodes) are connected by relations (or edges). A directed KG can formally be represented as a set of n triples: $\mathcal{T} = \{\langle h, r, t \rangle_i\}_i^n$ where each triple contains

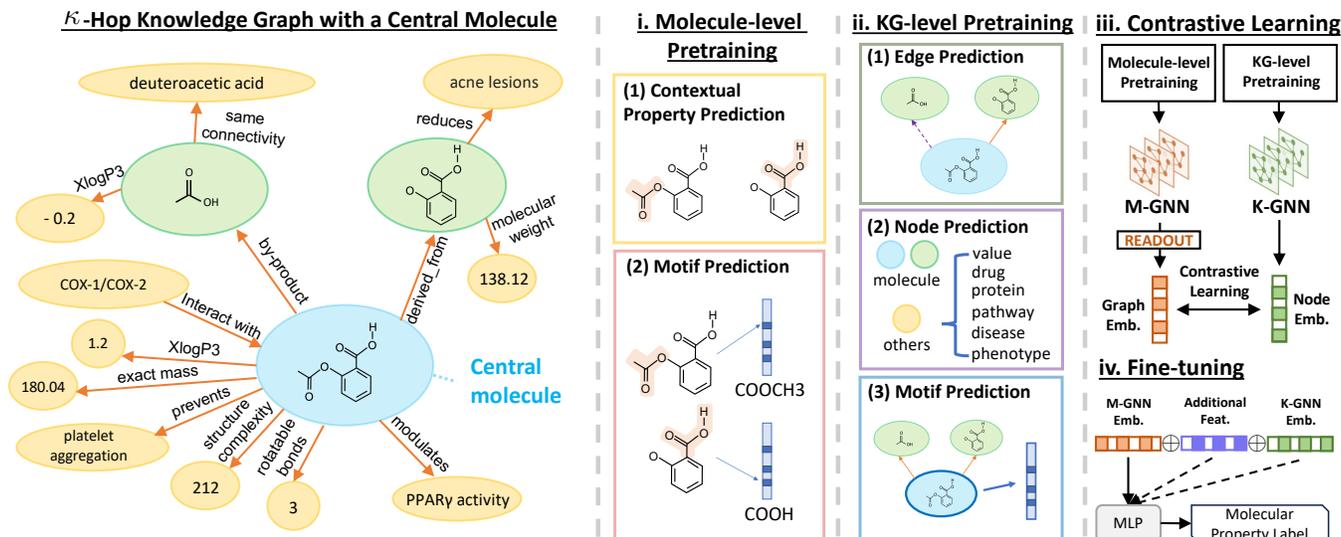


Figure 1: **Overview of our framework GODE.** *Left:* The κ -hop KG sub-graph consisting of molecule-relevant relational knowledge, originating from a central molecule. *Right:* We conduct (i) **Molecule-level Pre-training** on the molecular graphs with contextual property prediction and motif prediction tasks; (ii) **KG-level Pre-training** on the κ -hop KG sub-graphs of a central molecule with the tasks of edge prediction, node prediction, and motif prediction; (iii) **Contrastive Learning** to maximize the agreement between M-GNN and K-GNN, pre-trained by (i) and (ii), respectively; and (iv) **Fine-tuning** of our learned embedding, optionally enriched with extracted molecular-level features, for specific property predictions.

a head entity (h) and a tail entity (t), and a relation (r) connecting them. A KG G_k can also be viewed as a graph $G_k = (\mathcal{V}_k, \mathcal{E}_k)$ with a set of nodes \mathcal{V}_k and a set of edges \mathcal{E}_k .

Definition 3 (M-GNN) M-GNN is a graph encoder $f : \mathcal{M} \rightarrow \mathbb{R}^d$ encoding a MG to a vector \mathbf{h}_{MG} .

Definition 4 (K-GNN) K-GNN is a graph encoder $g : \mathcal{K} \rightarrow \mathbb{R}^d$ encoding the central molecule in a molecule KG sub-graph to a vector \mathbf{h}_{KG} .

Our GODE approach (illustrated in Figure 1) first conducts molecule-level pre-training to train an M-GNN and KG-level pre-training to train a K-GNN with a series of self-supervised tasks. Subsequently, we employ contrastive learning to enhance the alignment of molecule representations between the pre-trained M-GNN and K-GNN. Finally, we fine-tune our model for property prediction tasks.

Molecule-level Pre-training

Given a molecular graph $G_m = (\mathcal{V}_m, \mathcal{E}_m)$, we employ the GNN encoder to derive embeddings for atoms and bonds. To pre-train M-GNN, we employ two tasks described below.

(1) **Node-level Contextual Property Prediction.** We randomly select a node $v \in \mathcal{V}_m$ and its corresponding embedding \mathbf{h}_v . This embedding is then input into an output layer for predicting the contextual property. Contextual property prediction operates as a multi-class classification task. Here, the GNN’s output layer computes the probability distribution for potential contextual property labels linked to node v . These labels originate from the statistical attributes of the sub-graph centered on v (Rong et al. 2020).

(2) **Graph-level Motif Prediction.** The molecule graph embedding, represented as \mathbf{h}_{MG} , is also input into an output layer. This layer predicts the presence or absence of functional group motifs, which is detected by RD-Kit (Landrum et al. 2013). The embedding \mathbf{h}_{MG} is derived by applying mean pooling to all nodes: $\mathbf{h}_{MG} = \text{MEAN}(\mathbf{h}_{v_1}, \mathbf{h}_{v_2}, \dots, \mathbf{h}_{v_k} | v_1, v_2, \dots, v_k \in \mathcal{V}_m)$, where $\mathbf{h}_{v_1}, \mathbf{h}_{v_2}, \dots, \mathbf{h}_{v_k}$ are the learned node embeddings from the M-GNN’s final convolutional layer. This prediction task is a multi-label classification problem, where the GNN output layer predicts a binary label vector, indicating the presence or absence of each functional group motif in G_m .

During training, we employ a joint loss function, as shown in Eq. 1, to optimize both the node-level contextual property prediction and the graph-level motif prediction. This loss function encourages the M-GNN to accurately predict both the contextual properties of nodes and the functional group motifs’ presence or absence in MG.

$$\mathcal{L}_M = \sum_v \log P(p_v | \mathbf{h}_v) + \sum_{j=1}^n y_j \log P(M_j | \mathbf{h}_{MG}) + (1 - y_j) \log(1 - P(M_j | \mathbf{h}_{MG})), \quad (1)$$

where \mathcal{V}'_m is a set of randomly selected nodes; p_v is the contextual property label for the node v ; n is the number of all possible motifs; M_j is the presence of j -th motif.

After the molecule-level pre-training, M-GNN is able to encode a molecule to a vector \mathbf{h}_{MG} through mean pooling.

KG-level Pre-training

Embedding Initialization. Prior to the K-GNN pre-training, we use knowledge graph embedding (KGE) methods (Bordes et al. 2013; Yang et al. 2014; Sun et al. 2019; Balažević, Allen, and Hospedales 2019) to initialize the node and edge embeddings with entity and relation embeddings. KGE methods capture relational knowledge behind the structure and semantics of entities and relationships in the KG. The KGE model is trained on the entire KG (\mathcal{T}) and learns to represent each entity and relation as continuous vectors in a low-dimensional space. The resulting embedding vectors capture the semantic meanings and relationships between entities and relations. The loss functions of KGE methods depend on the scoring functions they use. For example, TransE (Bordes et al. 2013) learns embeddings for entities and relations in a KG by minimizing the difference between the sum of the head entity embedding (\mathbf{e}_h) and the relation embedding (\mathbf{r}_r), and the tail entity embedding (\mathbf{e}_t): $s(h, r, t) = -\|\mathbf{e}_h + \mathbf{r}_r - \mathbf{e}_t\|_p$, where $\|\cdot\|_p$ is the Lp norm. After training the KGE model, we obtain the entity embeddings \mathbf{e}_v and relation embeddings \mathbf{r}_e for each node v and edge e in the KG, providing a strong starting point.

Sub-graph Extraction. for the central molecule is a crucial step in KG-level pre-training. Inspired by the work of G-Meta (Huang and Zitnik 2020), we extract the sub-graph of each molecule to learn transferable knowledge from its surrounding nodes/edges in the biochemical KG. Specifically, for each central molecule, we extract a κ -hop sub-graph from the entire KG to capture its local neighborhood information. Given a molecule m_i , we first find its corresponding node v_i in the KG, $G_k = (\mathcal{V}_k, \mathcal{E}_k)$. We then iteratively extract a neighborhood sub-graph $\mathcal{N}_k(v_i, h)$ of depth h ($1 \leq h \leq \kappa$), centered at node v_i . The depth parameter h determines the number of edge traversals to include in the sub-graph. To avoid over-smoothing, we terminate the expansion of a graph branch upon reaching a non-molecule node. Formally, the sub-graph extraction process is defined as follows. Let $\mathcal{N}_k(v, 0)$ be a single node v . For $h > 0$, $\mathcal{N}_k(v, h)$ is defined recursively as:

$$\mathcal{N}_k(v, h) = \{v\} \cup \bigcup_{u \in \mathcal{N}_k(v, h-1)} \{u\} \cup \bigcup_{u \in \mathcal{M}} \{w : (u, w) \in \mathcal{E}_k\}, \quad (2)$$

where u denotes the set of neighboring nodes of v in the sub-graph $\mathcal{N}_k(v, h-1)$, and $w : (u, w) \in \mathcal{E}_k$ represents the set of nodes that share an edge with $u \in \mathcal{M}$ in the original KG G_k where \mathcal{M} is the set of molecule nodes. We define The κ -hop sub-graph for molecule m is given by $G_{\text{sub}(m, \kappa)} = (\mathcal{V}_{\text{sub}(m, \kappa)}, \mathcal{E}_{\text{sub}(m, \kappa)}) = \mathcal{N}_k(c, \kappa)$ where c is the corresponding node of m in $G_{\text{sub}(m, \kappa)}$.

The following loss function is used to pre-train K-GNN:

$$\begin{aligned} \mathcal{L}_K = & \lambda_m \underbrace{\sum_{j=1}^n \text{BCE}(y_j, P(M_j | \mathbf{h}_c))}_{\text{motif prediction}} + \lambda_n \underbrace{\text{CE}(v', P(v' | \mathbf{h}_v))}_{\text{node prediction}} \\ & + \lambda_e \underbrace{\text{CE}((u, v)', P((u, v)' | \mathbf{h}_u \oplus \mathbf{h}_v))}_{\text{edge prediction}} \end{aligned} \quad (3)$$

which includes three tasks shown in Figure 1 (ii):

- (1) **Edge Prediction**, a multi-class classification task aiming at correctly predicting the edge type between two nodes:
- (2) **Node Prediction**, a multi-class classification task predicting the category of a node in $G_{\text{sub}(m, \kappa)}$;
- (3) **Node-level Motif Prediction**, a multi-label classification task predicting the motif of the central molecule node c in $G_{\text{sub}(m, \kappa)}$. The motif labels are created by RDKit.

Here, $(u, v)'$ is the label of edge between the nodes u and v . v' is the label of node v , \oplus denotes the embedding concatenation. y_j is binary indicator, $P(M_j | \mathbf{h}_c)$ is the predicted probability of central molecule c has the j -th functional group motif M_j given its embedding \mathbf{h}_c . λ_e , λ_m , and λ_n are balancing hyperparameters.

After the KG-level pre-training, K-GNN can encode a molecule to a vector \mathbf{h}_{KG} given its surrounding nodes.

Contrastive Learning

Inspired by the success of previous works (Radford et al. 2021; Seidl et al. 2023; Sanchez-Fernandez et al. 2022) that apply contrastive learning to transfer knowledge across different modalities, we follow their steps using InfoNCE as the loss function to conduct contrastive learning between molecule graph and KG sub-graph. We construct the training set $\mathcal{D} = \mathcal{D}^+ \cup \mathcal{D}^- = \{(m_i, s_i), y_i\}_N$, where $\mathcal{D}^+ = \{(m_i, G_{\text{sub}(m_i, \kappa)}), y_i = 1\}_{N_p}$ is a set of positive samples and $\mathcal{D}^- = \{(m_i, G_{\text{sub}(m_j, \kappa)})_{j \neq i}, y_i = 0\}_{N-N_p}$ is a set of negative samples. To make the task more challenging, we further divide \mathcal{D}^- into \mathcal{D}_{rand}^- and \mathcal{D}_{nbr}^- , which are (1) randomly sampled from all negative molecule-centric KG sub-graphs, and (2) sampled from the sub-graphs of the neighbor molecule nodes connected to the positive molecule node, respectively. The loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{InfoNCE}} = & -\frac{1}{N} \sum_{i=1}^N \left[y_i \log(\text{sim}(f(m_i), g(s_i))) \right. \\ & \left. + (1 - y_i) \log(1 - \text{sim}(f(m_i), g(s_i))) \right], \end{aligned} \quad (4)$$

where $\text{sim}(f(m_i), g(s_i)) = \frac{\exp(\tau^{-1} \mathbf{h}_{\text{MG}(i)}^T \mathbf{h}_{\text{KG}(i)})}{\exp(\tau^{-1} \mathbf{h}_{\text{MG}(i)}^T \mathbf{h}_{\text{KG}(i)}) + 1}$, y_i is the binary label, m_i and s_i are the paired MG and KG sub-graph in the training data, τ^{-1} is the inverse temperature.

Fine-tuning for Downstream Tasks

Upon completing molecule- and KG-level pre-training combined with contrastive learning, we obtain two GNN encoders, f and g , which respectively encode molecules and KG sub-graphs into vectors. Following previous works (Rong et al. 2020; Fang et al. 2023; Wu et al. 2018; Yang et al. 2019), we employ RDKit to extract additional molecule-level features \mathbf{h}_f . A joint representation is formed by $\mathbf{h}_{\text{joint}} = \mathbf{h}_{\text{MG}} \oplus \mathbf{h}_f \oplus \mathbf{h}_{\text{KG}}$, with \oplus representing concatenation. This representation is then utilized to predict the target property y using a multi-layer perceptron (MLP) with an activation function. For multi-label classification, we use Binary Cross-Entropy (BCE) loss with sigmoid activation, and for regression, we use Mean Squared Error (MSE) loss.

Triples: 2,523,867 # Entities: 184,819 # Relations: 39 # Entity Types: 7 # Molecules: 65,454

Entity Types

molecule, gene/protein, disease, effect/phenotype, drug, pathway, value

Relations

drug_protein, contraindication, indication, off-label use, drug_drug, drug_effect, defined_bond_stereo_count, tpsa, rotatable_bond_count, xlogp3-aa, structure_complexity, covalent_unit_count, defined_atom_stereo_count, molecular_weight, hydrogen_bond_donor_count, undefined_bond_stereo_count, isotope_atom_count, exact_mass, mono_isotopic_weight, total_formal_charge, hydrogen_bond_acceptor_count, non-hydrogen_atom_count, tautomer_count, undefined_atom_stereo_count, xlogp3, cooccurrence_molecule_molecule, cooccurrence_molecule_disease, cooccurrence_molecule_gene/protein, neighbor_2d, neighbor_3d, has_same_connectivity, has_component, has_isotopologue, has_parent, has_stereoisomer, to_drug, closematch, type, in_pathway

Table 1: Overview of MolKG, a biochemical dataset we construct from PubChemRDF and PrimeKG.

Experiments

Experimental Setting

Data Sources. (1) Molecule-level pre-training data: The pre-training data for our molecule-level M-GNN is derived from the same unlabelled dataset of 11 million molecules utilized by GROVER. This dataset encompasses sources such as ZINC15 (Sterling and Irwin 2015) and ChEMBL (Gaulton et al. 2012). We randomly split this dataset into two subsets with a 9:1 ratio for training and validation. (2) Knowledge graph-level pre-training data: For the KG-level pre-training, we retrieve KG triples related to the molecules from PubChemRDF and PrimeKG. These include various subdomains and properties from PubChemRDF, as well as 3-hop sub-graphs for all 7957 drugs from PrimeKG. We show an overview of the dataset (MolKG) in Table 1. The dataset is divided into training and validation sets with a 9:1 ratio. *The construction details of the dataset are placed in Appendix.* (3) Contrastive learning data: we set the negative sampling ratio as $|\mathcal{D}^-|/|\mathcal{D}^+| = 32$ and retain a 1 : 1 ratio for $\mathcal{D}_{rand}^- : \mathcal{D}_{nbr}^-$. Training and validation samples are in a 0.95 : 0.05 ratio. (4) Downstream task datasets: The effectiveness of our model is tested utilizing the comprehensive MoleculeNet dataset (Wu et al. 2018; Huang et al. 2021)¹, which contains 6 classification and 5 regression datasets for molecular property prediction. We place detailed descriptions of these datasets in the Appendix. To fine-tune the model, we calculate the mean and standard deviation of the ROC-AUC for classification tasks and RMSE/MAE for regression tasks. Scaffold splitting with three random seeds was employed with a training/validation/testing ratio of 8:1:1 across all datasets, aligning with previous studies (Rong et al. 2020; Fang et al. 2023).

Baselines. We compare our proposed model with several popular baselines in molecular property prediction tasks, which include GCN (Kipf and Welling 2016), GIN (Xu et al. 2018), SchNet (Schütt et al. 2017), MPNN (Gilmer et al. 2017), DMPNN (Yang et al. 2019), MGCN (Lu et al. 2019), N-GRAM (Liu, Demirel, and Liang 2019), Hu et al (Hu et al. 2019), GROVER (Rong et al. 2020), MGSSL (Zhang et al. 2021), KGE_NFM (Ye et al. 2021) with our MolKG, MolCLR (Wang et al. 2021b), and KANO (Fang et al. 2023).

Implementation. For molecule-level pre-training, we employ GROVER (Rong et al. 2020), and for KG-level pre-

training, we utilize GINE (Hu et al. 2019). TransE initializes the KG embeddings over a span of 10 epochs. Our settings include $\lambda_e = 1.5$, $\lambda_m = 1.8$, and $\lambda_n = 1.5$. Both M-GNN and K-GNN have a hidden size of 1,200. We adopt a temperature $\tau = 1.0$ for contrastive learning. Early stopping is anchored to validation loss. During fine-tuning, embeddings from K-GNN remain fixed, updating only the parameters of M-GNN. We use Adam optimizer with the Noam learning rate scheduler (Vaswani et al. 2017). All tests are performed with two AMD EPYC 7513 32-core Processors, 528GB RAM, 8 NVIDIA A6000 GPUs, and CUDA 11.7. *More implementation details and the hyper-parameter study are placed in Appendix.*

Results on Molecule Property Prediction

Table 2 presents comparative performance metrics for classification and regression tasks, respectively. It is clear from the data that our proposed method, GODE, consistently outperforms the baseline models in most tasks. Specifically, in classification tasks, GODE achieves SOTA results across all tasks. Amongst the competitors, KANO stands out, consistently showcasing performance close to our method. Intriguingly, KANO, as a knowledge-driven model, augments molecular structures by integrating information about chemical elements from its ElementKG. This underlines the substantial advantage of leveraging external knowledge in predicting molecular properties. On the regression front, GODE attains best results in 4 out of 5 tasks. This consistent high performance, irrespective of the nature of the task, underscores our model’s adaptability and reliability. Cumulatively, our approach yields a relative improvement of 23.6% across all tasks (12.7% for classification and 34.4% for regression tasks). When compared with the SOTA model, KANO, GODE records improvements of 2.2% and 7.2% for classification and regression tasks, respectively.

To analyze the effects of GODE’s variants, we conduct ablation studies in Figure 2, which are discussed as follows.

Effect of the Integration of MolKG. To assess the impact of integrating our molecule-centric KG - MolKG, into molecule property prediction, we compare Case 8 in Figure 2 with our backbone M-GNN model, GROVER. Specifically, Case 8 melds GROVER ($\mathbf{h}_{MG} \oplus \mathbf{h}_f$) with the static KG embedding (\mathbf{h}_{KGE}), which is trained using the KGE method. Our observations indicate that infusing the KG boosts performance across all tasks, resulting in a noteworthy 14.3% overall enhancement. Moreover, when all variants of GODE

¹<https://moleculenet.org/datasets-1>

Dataset	Classification (Higher is Better)						Regression (Lower is Better)				
	BBBP	SIDER	ClinTox	BACE	Tox21	ToxCast	FreeSolv	ESOL	Lipophilicity	QM7	QM8
# Molecules	2039	1427	1478	1513	7831	8575	642	1128	4200	6830	21786
# Tasks	1	27	2	1	12	617	1	1	1	1	12
GCN	71.8 ± 0.9	53.6 ± 0.3	62.5 ± 2.8	71.6 ± 2.0	70.9 ± 0.3	65.0 ± 6.1	2.870 ± 0.140	1.430 ± 0.050	0.712 ± 0.049	122.9 ± 2.2	0.037 ± 0.001
GIN	65.8 ± 4.5	57.3 ± 1.6	58.0 ± 4.4	70.1 ± 5.4	74.0 ± 0.8	66.7 ± 1.5	2.765 ± 0.180	1.452 ± 0.020	0.850 ± 0.071	124.8 ± 0.7	0.037 ± 0.001
SchNet	84.8 ± 2.2	54.5 ± 3.8	71.7 ± 4.2	76.6 ± 1.1	76.6 ± 2.5	67.9 ± 2.1	3.215 ± 0.755	1.045 ± 0.064	0.909 ± 0.098	74.2 ± 6.0	0.020 ± 0.002
MPNN	91.3 ± 4.1	59.5 ± 3.0	87.9 ± 5.4	81.5 ± 4.4	80.8 ± 2.4	69.1 ± 1.3	1.621 ± 0.952	1.167 ± 0.430	0.672 ± 0.051	111.4 ± 0.9	0.015 ± 0.001
DMPNN	91.9 ± 3.0	63.2 ± 2.3	89.7 ± 4.0	85.2 ± 5.3	82.6 ± 2.3	71.8 ± 1.1	1.673 ± 0.082	1.050 ± 0.008	0.683 ± 0.016	103.5 ± 8.6	0.016 ± 0.001
MGCN	85.0 ± 6.4	55.2 ± 1.8	63.4 ± 4.2	73.4 ± 3.0	70.7 ± 1.6	66.3 ± 0.9	3.349 ± 0.097	1.266 ± 0.147	1.113 ± 0.041	77.6 ± 4.7	0.022 ± 0.002
N-GRAM	91.2 ± 1.3	63.2 ± 0.5	85.5 ± 3.7	87.6 ± 3.5	76.9 ± 2.7	-	2.512 ± 0.190	1.100 ± 0.160	0.876 ± 0.033	125.6 ± 1.5	0.032 ± 0.003
HU. et.al	70.8 ± 1.5	62.7 ± 0.8	72.6 ± 1.5	84.5 ± 0.7	78.7 ± 0.4	65.7 ± 0.6	2.764 ± 0.002	1.100 ± 0.006	0.739 ± 0.003	113.2 ± 0.6	0.022 ± 0.001
GROVER _{Large, GTrans}	86.2 ± 3.9	57.6 ± 1.6	74.7 ± 4.4	82.5 ± 4.4	76.9 ± 2.3	66.7 ± 2.6	2.445 ± 0.761	1.028 ± 0.145	0.890 ± 0.050	95.3 ± 5.6	0.020 ± 0.003
MGSSL	70.5 ± 1.1	64.1 ± 0.7	80.7 ± 2.1	79.7 ± 0.8	76.4 ± 0.4	64.1 ± 0.7	-	-	-	-	-
MolCLR	73.3 ± 1.0	61.2 ± 3.6	89.8 ± 2.7	82.8 ± 0.7	74.1 ± 5.3	65.9 ± 2.1	2.301 ± 0.247	1.113 ± 0.023	0.789 ± 0.009	90.0 ± 1.7	0.019 ± 0.013
MolCLRGTrans	76.7 ± 2.2	63.3 ± 2.5	89.3 ± 3.1	87.7 ± 1.8	80.2 ± 3.2	70.4 ± 2.1	2.124 ± 0.223	0.982 ± 0.109	0.767 ± 0.064	88.9 ± 4.8	0.018 ± 0.012
KGE _{NFM} _{w/ MolKG}	92.4 ± 2.4	65.3 ± 1.4	87.3 ± 2.0	78.1 ± 2.1	79.8 ± 3.3	72.6 ± 1.8	1.942 ± 0.441	1.027 ± 0.201	0.877 ± 0.071	87.6 ± 3.2	0.016 ± 0.001
KANO _{CMPNN}	92.6 ± 1.8	65.5 ± 1.6	92.9 ± 1.1	90.7 ± 3.1	81.8 ± 1.1	72.5 ± 1.9	1.320 ± 0.244	0.902 ± 0.104	0.641 ± 0.012	66.5 ± 3.7	0.013 ± 0.001
KANO _{GTrans}	93.7 ± 2.3	63.8 ± 1.2	93.6 ± 0.7	90.4 ± 1.5	81.2 ± 1.8	72.5 ± 1.5	1.443 ± 0.315	0.914 ± 0.092	0.651 ± 0.018	63.6 ± 4.1	0.013 ± 0.002
GODE (ours)	94.8 ± 1.9	67.4 ± 1.4	94.7 ± 2.9	92.0 ± 2.2	84.3 ± 1.2	73.4 ± 0.9	1.048 ± 0.314	0.746 ± 0.128	0.743 ± 0.043	57.2 ± 3.0	0.013 ± 0.001

Table 2: Performance on six classification benchmarks (ROC-AUC; higher is better) and five regression benchmarks (RMSE for FreeSolv, ESOL, and Lipophilicity; MAE for QM7/QM8; lower is better). We report the mean and standard deviation. The Top-3 results are highlighted in bold. The backbone model is shaded in grey, and models utilizing the backbone are shaded in yellow. The table is divided into three sections: non-KG methods, other KG-based methods, and our method.

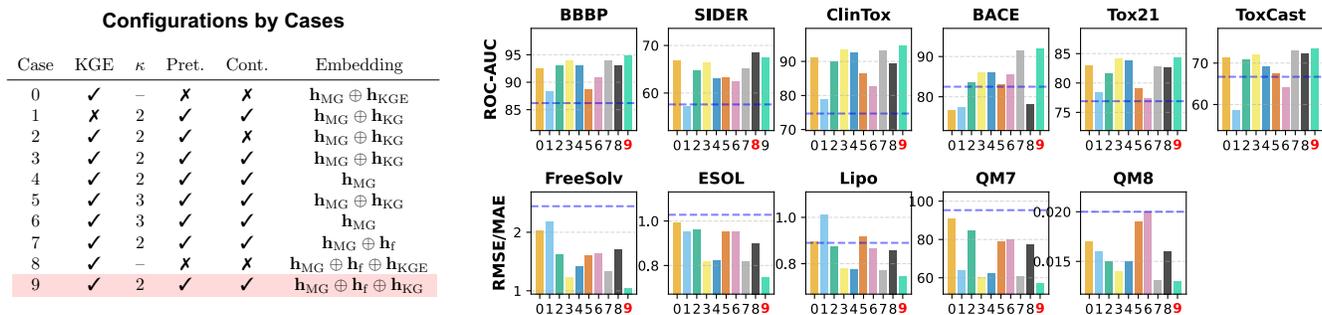


Figure 2: Ablation study configurations and results. (Left) Configurations. “KGE”: KG embedding initialization. “ κ ”: κ -hop KG subgraph. “Pret.”: KG-level pre-training. “Cont.”: contrastive learning. “Embedding”: input to MLP for fine-tuning. (Right) Performance comparison across different datasets and configurations. We highlight the best configuration for each dataset in red. The dotted blue lines denote the performance achieved by the backbone model (GROVER).

are deployed (as in Case 9), a significant uplift of 23.2% in performance over GROVER is realized.

Effect of KG-level Pre-training and Contrastive Learning. Through a side-by-side comparison of Cases 0, 2, and 3 in Figure 2, we discern the value of K-GNN pre-training and contrastive learning. Standalone K-GNN pre-training (Case 2) yields a modest boost of 4.5%, with a particularly slight edge in classification tasks at 0.1%. However, when paired with contrastive learning and leveraging both h_{MG} and h_{KG} for fine-tuning, as in Case 3, the surge in performance is notable, reaching an overall enhancement of 13.6% over the baseline Case 0. A testament to the effectiveness of this approach can be seen in the BBBP dataset. The molecule acetylsalicylate, better known as aspirin, posed a prediction challenge to both our M-GNN model and the methods in Cases 0 and 2. Yet, when Case 3 employed relational knowledge from its KG sub-graph (e.g., [acetylsalicylate, indication, neurological conditions]) alongside contrastive learning, it managed to make accurate predictions. This example

underscores the pivotal role of contrastive learning in refining molecular property predictions.

Efficacy of Knowledge Transfer. The influence of contrastive learning in transferring domain knowledge from the biochemical KG to the molecular representation h_{MG} is discerned by examining Cases 3, 4, 5, 6, and contrasting GROVER (backbone) with Cases 7 and 9. Notably, while the M-GNN embeddings of GODE (represented by Cases 4 and 6) do not quite surpass the bi-level concatenated embeddings (Cases 3 and 5), they come notably close. More compelling is Case 7, which parallels Case 9 and outperforms GROVER by a striking 21.0% (with 12.0% in classification and 30.1% in regression). The distinguishing feature of Case 7 that provides an edge over GROVER is its enriched h_{MG} , an enhancement absent in GROVER. This underscores GODE’s prowess in biochemical knowledge transfer to molecular representations.

Mutual Benefit of Bi-level Self-supervised Pre-training. In addition to the insights provided in Figure 2, we con-

(a) Effect of Pre-training on M-GNN and K-GNN												
M-GNN Pret.	K-GNN Pret.	BBBP	SIDER	ClinTox	BACE	Tox21	ToxCast	FreeSolv	ESOL	Lipo	QM7	QM8
✓	✓	94.8	67.4	94.7	92.0	84.3	73.4	1.048	0.746	0.743	57.2	0.013
×	✓	92.2	62.6	89.4	89.8	80.6	70.8	1.313	0.834	0.708	64.6	0.016
✓	×	93.2	66.7	90.7	81.6	83.1	71.9	1.563	0.841	0.876	74.4	0.017
×	×	88.9	62.1	88.4	84.1	81.6	69.4	1.944	0.978	0.845	77.9	0.017

(b) Effect of Relationship Exclusion from MolKG												
Knowledge Graph	BBBP	SIDER	ClinTox	BACE	Tox21	ToxCast	FreeSolv	ESOL	Lipo	QM7	QM8	
MolKG	94.8	67.4	94.7	92.0	84.3	73.4	1.048	0.746	0.743	57.2	0.013	
w/o indication	93.8	65.7	93.4	91.6	84.2	73.0	1.063	0.754	0.751	58.1	0.013	
w/o <i>xlogp3</i> & <i>xlogp3-aa</i>	93.7	66.0	94.2	91.1	83.0	72.8	1.189	0.789	0.782	57.8	0.012	
w/o <i>tautomer_cnt</i> & <i>covalent_unit_cnt</i>	94.3	66.5	93.1	90.9	83.5	72.5	1.272	0.761	0.759	61.7	0.014	
w/o <i>nbr_2d</i> & <i>nbr_3d</i> & <i>has_same_conn</i>	95.0	67.3	93.6	91.3	84.3	72.7	1.058	0.749	0.748	57.6	0.013	

Table 3: Study the effects of (top) bi-level self-supervised pre-training and (below) relationship exclusion from MolKG.

ducted an in-depth analysis of the impact of pre-training M-GNN and K-GNN on the performance of GODE, as detailed in Table 3(a). The findings clearly underscore the mutual benefit of both M-GNN and K-GNN pre-training to the efficacy of our framework. Notably, an improved performance on the Lipophilicity dataset was observed when the M-GNN pre-training was omitted, presenting an intriguing aspect for further investigation.

Impact of Relationship Exclusion. We investigated the impact of removing specific relationships from MolKG on both classification and regression datasets (in Table 3(b)). For classification, excluding “tautomer_count” and “covalent_unit_count” led to the largest performance drop on ClinTox, while removing structural similarity relationships slightly improved results on BBBP. For regression, removing “xlogp3” and “xlogp3-aa” substantially increased the error on solvation and lipophilicity predictions, aligning with the physical meaning of these features. Removing “tautomer_count” and “covalent_unit_count” also notably impacted predicting solvation and quantum properties. This analysis reveals the variable significance of different relationships, with the most consistent impact observed for “xlogp3” and “xlogp3-aa” on solvation and lipophilicity tasks.

Embedding Visualization. In the t-SNE visualization presented in Figure 3, the GROVER embeddings highlight molecules from varying scaffolds intermingling, signaling a significant avenue for refinement. Particularly in the Tox21 task, these embeddings appear sparse. When enhanced with KANO, there is a noticeable delineation of clusters, reflecting the constructive influence of integrating external knowledge into molecular representations. Nonetheless, a residual overlap of molecules from different scaffolds still persists. Progressing to the GODE visualization, the clusters exhibit further refinement, achieving pronounced distinctiveness with minimal scaffold overlap, outperforming KANO, and securing the lowest Davies–Bouldin Index (DBI), which signifies the effectiveness of GODE.

Conclusion

We introduced GODE, a framework that enhances molecule representations through bi-level self-supervised pre-training

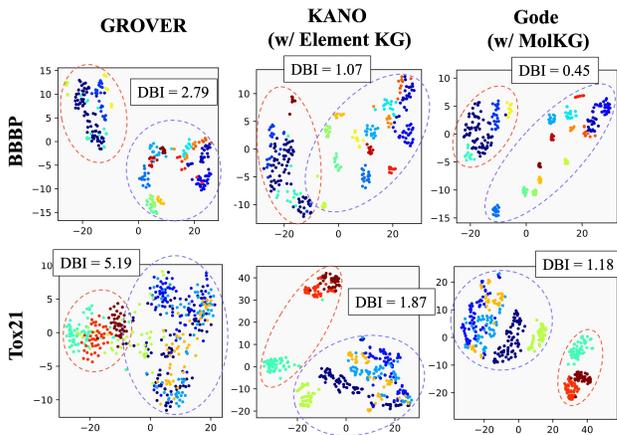


Figure 3: t-SNE visualization of molecule embeddings across two tasks. Each color represents a unique scaffold (molecule substructure). We compare the embeddings from GROVER, KANO, and GODE. The clustering quality is assessed using the DB index.

and contrastive learning, leveraging biochemical domain knowledge. Our empirical results demonstrate its effectiveness in molecular property prediction tasks. Future work will focus on expanding the coverage of MolKG and identifying crucial knowledge elements for optimizing molecular representations. This research lays groundwork for advancements in drug discovery applications.

Acknowledgments

Research was supported in part by US DARPA INCAS Program No. HR0011-21-C0165 and BRIES Program No. HR0011-24-3-0325, National Science Foundation IIS-19-56151, the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329.

References

- Argelaguet, R.; Arnol, D.; Bredikhin, D.; Deloro, Y.; Velten, B.; Marioni, J. C.; and Stegle, O. 2020. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21(1): 1–17.
- Balažević, I.; Allen, C.; and Hospedales, T. M. 2019. Tucker: Tensor factorization for knowledge graph completion. *arXiv preprint arXiv:1901.09590*.
- Bécigneul, G.; Ganea, O.-E.; Chen, B.; Barzilay, R.; and Jaakkola, T. S. 2020. Optimal transport graph neural networks.
- Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl.1): D267–D270.
- Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.
- Chandak, P.; Huang, K.; and Zitnik, M. 2023a. Building a knowledge graph to enable precision medicine. *Nature Scientific Data*.
- Chandak, P.; Huang, K.; and Zitnik, M. 2023b. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1): 67.
- Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28.
- Fang, Y.; Zhang, Q.; Zhang, N.; Chen, Z.; Zhuang, X.; Shao, X.; Fan, X.; and Chen, H. 2023. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 1–12.
- Fu, G.; Batchelor, C.; Dumontier, M.; Hastings, J.; Willighagen, E.; and Bolton, E. 2015. PubChemRDF: towards the semantic annotation of PubChem compound and substance databases. *Journal of cheminformatics*, 7(1): 1–15.
- Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1): D1100–D1107.
- Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272. PMLR.
- Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; and Head-Gordon, T. 2020. Learning to make chemical predictions: the interplay of feature representation, data, and machine learning methods. *Chem*, 6(7): 1527–1542.
- Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2019. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*.
- Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y. H.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; and Zitnik, M. 2021. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
- Huang, K.; and Zitnik, M. 2020. Graph meta learning via local subgraphs. *Advances in neural information processing systems*, 33: 5862–5874.
- Jaeger, S.; Fulle, S.; and Turk, S. 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58(1): 27–35.
- Jin, W.; Coley, C. W.; Barzilay, R.; and Jaakkola, T. 2017. Predicting organic reaction outcomes with weisfeiler-lehman network. *arXiv preprint arXiv:1709.04555*.
- Jin, W.; Yang, K.; Barzilay, R.; and Jaakkola, T. 2018. Learning multimodal graph-to-graph translation for molecular optimization. *arXiv preprint arXiv:1812.01070*.
- Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Landrum, G.; et al. 2013. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8.
- Lee, H.; Ahn, S.; Seo, S.-W.; Song, Y. Y.; Yang, E.; Hwang, S.-J.; and Shin, J. 2021. RetCL: A Selection-based Approach for Retrosynthesis via Contrastive Learning. *arXiv:2105.00795*.
- Li, M.; Zhou, J.; Hu, J.; Fan, W.; Zhang, Y.; Gu, Y.; and Karypis, G. 2021. Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. *ACS omega*, 6(41): 27233–27238.
- Li, M. M.; Huang, K.; and Zitnik, M. 2022. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 1–17.
- Liao, R.; Zhao, Z.; Urtasun, R.; and Zemel, R. S. 2019. Lanczosnet: Multi-scale deep graph convolutional networks. *arXiv preprint arXiv:1901.01484*.
- Liu, S.; Demirel, M. F.; and Liang, Y. 2019. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems*, 32.
- Lu, C.; Liu, Q.; Wang, C.; Huang, Z.; Lin, P.; and He, L. 2019. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1052–1060.
- Mansimov, E.; Mahmood, O.; Kang, S.; and Cho, K. 2019. Molecular geometry prediction using a deep generative graph neural network. *Scientific reports*, 9(1): 20381.
- Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

- Rogers, D.; and Hahn, M. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5): 742–754.
- Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; and Huang, J. 2020. Self-Supervised Graph Transformer on Large-Scale Molecular Data. *arXiv:2007.02835*.
- Sanchez-Fernandez, A.; Rumetshofer, E.; Hochreiter, S.; and Klambauer, G. 2022. Contrastive learning of image- and structure-based representations in drug discovery. In *ICLR2022 Machine Learning for Drug Discovery*.
- Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; and Müller, K.-R. 2017. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30.
- Seidl, P.; Renz, P.; Dyubankova, N.; Neves, P.; Verhoeven, J.; Wegner, J. K.; Segler, M.; Hochreiter, S.; and Klambauer, G. 2022. Improving few- and zero-shot reaction template prediction using modern hopfield networks. *Journal of chemical information and modeling*, 62(9): 2111–2120.
- Seidl, P.; Vall, A.; Hochreiter, S.; and Klambauer, G. 2023. Enhancing activity prediction models in drug discovery with the ability to understand human language. *arXiv preprint arXiv:2303.03363*.
- Shang, C.; Liu, Q.; Chen, K.-S.; Sun, J.; Lu, J.; Yi, J.; and Bi, J. 2018. Edge attention-based multi-relational graph convolutional networks. *arXiv preprint arXiv: 1802.04944*.
- Somnath, V. R.; Bunne, C.; and Krause, A. 2021. Multi-scale representation learning on proteins. *Advances in Neural Information Processing Systems*, 34: 25244–25255.
- Stärk, H.; Beaini, D.; Corso, G.; Tossou, P.; Dallago, C.; Günnemann, S.; and Lió, P. 2022. 3D Infomax improves GNNs for Molecular Property Prediction. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 20479–20502. PMLR.
- Sterling, T.; and Irwin, J. J. 2015. ZINC 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11): 2324–2337.
- Su, B.; Du, D.; Yang, Z.; Zhou, Y.; Li, J.; Rao, A.; Sun, H.; Lu, Z.; and Wen, J.-R. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*.
- Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.
- Tong, T.; Gray, K.; Gao, Q.; Chen, L.; Rueckert, D.; Initiative, A. D. N.; et al. 2017. Multi-modal classification of Alzheimer’s disease using nonlinear graph fusion. *Pattern recognition*, 63: 171–181.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972): 47–60.
- Wang, H.; Li, W.; Jin, X.; Cho, K.; Ji, H.; Han, J.; and Burke, M. D. 2021a. Chemical-reaction-aware molecule representation learning. *arXiv preprint arXiv:2109.09888*.
- Wang, X.; Li, Z.; Jiang, M.; Wang, S.; Zhang, S.; and Wei, Z. 2019. Molecule property prediction based on spatial graph embedding. *Journal of chemical information and modeling*, 59(9): 3817–3828.
- Wang, Y.; Wang, J.; Cao, Z.; and Farimani, A. 2021b. MolCLR: Molecular contrastive learning of representations via graph neural networks. *arXiv 2021. arXiv preprint arXiv:2102.10056*.
- Wang, Y.; Zhao, Y.; Shah, N.; and Derr, T. 2022. Imbalanced graph classification via graph-of-graph neural networks. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2067–2076.
- Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; and Pande, V. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2): 513–530.
- Xu, B.; Lu, Y.; Li, C.; Yue, L.; Wang, X.; Hao, N.; Fu, T.; and Chen, J. 2024. SMILES-Mamba: Chemical Mamba Foundation Models for Drug ADMET Prediction. *arXiv preprint arXiv:2408.05696*.
- Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations*.
- Xu, Y.; Pei, J.; and Lai, L. 2017. Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *Journal of chemical information and modeling*, 57(11): 2672–2685.
- Yang, B.; Yih, W.-t.; He, X.; Gao, J.; and Deng, L. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.
- Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8): 3370–3388.
- Ye, Q.; Hsieh, C.-Y.; Yang, Z.; Kang, Y.; Chen, J.; Cao, D.; He, S.; and Hou, T. 2021. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nature communications*, 12(1): 6775.
- Yuan, H.; and Ji, S. 2020. Structpool: Structured graph pooling via conditional random fields. In *Proceedings of the 8th International Conference on Learning Representations*.
- Zhang, Z.; Liu, Q.; Wang, H.; Lu, C.; and Lee, C.-K. 2021. Motif-based Graph Self-Supervised Learning for Molecular Property Prediction. *arXiv:2110.00987*.
- Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; and Yang, Y. 2019. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of Chemical Information and Modeling*, 60(1): 47–55.