Bi-level Contrastive Learning for Knowledge-Enhanced Molecule Representations

Pengcheng Jiang*, Cao Xiao†, Tianfan Fu*, Jimeng Sun*

*University of Illinois at Urbana Champaign, †GE Healthcare
*{pj20, tianfanf, jimeng}@illinois.edu, †danicaxiao@gmail.com

Abstract

Molecule representation learning underpins diverse downstream applications such as molecular property and side effect understanding and prediction. In this paper, we recognize the two-level structure of individual molecule as having intrinsic graph structure as well as being a node in a large molecule knowledge graph, and present GODE, a new approach that seamlessly integrates graph representations of individual molecules with multi-domain biomedical data from knowledge graphs. By pre-training two graph neural networks (GNNs) on different graph structures, combined with contrastive learning, GODE adeptly fuses molecular structures with their corresponding knowledge graph substructures. This fusion results in a more robust and informative representation, enhancing molecular property prediction by harnessing both chemical and biological information. Fine-tuned on 11 chemical property tasks, our model surpasses benchmarks, achieving an average ROC-AUC improvement of 14.5%, 9.8%, and 7.3% on BBBP, SIDER, and Tox21 datasets. In regression tasks on ESOL and QM7 datasets, we achieve average improvements of 21.0% and 29.6% improvements in RMSE and MAE, setting a new field benchmark.

Introduction

Recent years have witnessed a surge of efforts in tailoring machine learning models for chemical and biological data (Wang et al. 2021a; Li, Huang, and Zitnik 2022; Somnath, Bunne, and Krause 2021; Wang et al. 2023). Within this realm, a pivotal concern revolves around formulating potent representations for intricate molecular structures, essential for subsequent tasks (Yang et al. 2019; Haghighatlari et al. 2020). To address this, the ascendancy of graph neural networks (GNNs) has been notable, adeptly leveraging the inherent graph-like essence of the data to facilitate representation learning (Li et al. 2021; Hu et al. 2019). Nevertheless, the conventional practice of employing molecular graphs as GNN input can inadvertently truncate their potential for effective and robust representation.

Molecular data (e.g., chemical and biological datasets) inherently manifest multimodality (Tong et al. 2017; Argelaguet et al. 2020). For individual molecules, their molecular structures naturally lend themselves to graph representations, with atoms as nodes and bonds as edges, while for sets of molecules, their relations are also described in knowledge graphs (KGs), where each molecule is a node. For example, such KGs include UMLS (Bodenreider 2004), PrimeKG (Chandak, Huang, and Zitnik 2023a), and PubChemRDF (Fu et al. 2015). We hypothesize that if we could properly fuse the two types of graph data: the molecular graphs and the molecule-centric KG sub-graphs across molecules, we could learn more enriched molecule representation that would exhibit more accurate and robust prediction performance.

Previous attempts have sought to unify molecule structures with knowledge graphs for property prediction. For instance, Ye et al. (2021) combines molecule embeddings with static KG embeddings (Bordes et al. 2013). However, such amalgamations sometimes fail to capture the local information of molecules in the KG, resulting in marginal prediction enhancements. On the other hand, Fang et al. (2022, 2023) highlight the benefits of improving molecule representations using contrastive learning, supported by their designed chemical element KG. This approach results in more visible performance improvements, showing the value of using KGs with molecular data. Our work aims to find new ways to integrate biomedical knowledge graphs into molecular prediction models.

In this study, we propose a new approach, coined as "**G**raph as a N**ode**" (GODE), designed to pre-train Graph Neural Networks (GNNs). Our approach encompasses bi-level self-supervised tasks, targeting both molecular structures and their corresponding sub-graphs within the knowledge graph (KG). By synergizing this strategy with contrastive learning, (GODE) yields more robust embeddings for molecule property predictions.

Our major contributions can be summarized as follows:

- **A new paradigm to connect knowledge and data**. Our GODE method offers a new paradigm for the integration between molecular structures with their corresponding knowledge graphs, which not only yields richer and better molecular representation in our use case but also can be extended to other application domains.
- **Robust embedding enhancement**. For molecular representation, they need to be robust to ensure accurate and consistent predictions across diverse molecular datasets. By integrating information from different domains for the same molecule, our approach leverages the shared knowledge across modalities, ensuring a more comprehensive representation. By employing bi-level self-

supervised pre-training with contrastive learning, we significantly enhance the robustness and reliability of the embeddings. This ensures that the generated embeddings are more adept at predicting molecular properties, providing a solid foundation for various applications.

- **A new molecular knowledge graph MolKG**. We have constructed MolKG, a comprehensive knowledge graph tailored specifically for molecular data. MolKG encapsulates vast molecular information and facilitates enhanced knowledge-driven molecular analyses.

To evaluate the performance of GODE, we conducted extensive experiments across 11 chemical property prediction tasks. We compared GODE to state-of-the-art methods such as GROVER (Rong et al. 2020), MolCLR (Wang et al. 2021a), and KANO (Fang et al. 2023). Our evaluations demonstrate Gode's superior performance in molecular property prediction, surpassing the baselines by 10.3% and 23.2% for classification and regression tasks, respectively.

Related Works

Graph-based Molecular Representation Learning. Over the years, various streams of molecular representation methods have been proposed. They encompass traditional fingerprint-based approaches (Rogers and Hahn 2010; Jaeger, Fulle, and Turk 2018) and modern graph neural network (GNN) methods (Jin et al. 2017; Coley et al. 2019; Jin et al. 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 subgraphs. This advantage yields enhanced representations for chemical nodes, bonds, and entire molecules (Cai et al. 2022; 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 knowledge graphs (KGs) have emerged to capture interconnections among diverse entities like genes, proteins, diseases, and drugs (Belleau et al. 2008; Szklarczyk et al. 2019; Piñero et al. 2020; 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 chemical representation learning frameworks. Three main aspects of the molecular property attract researchers: quantum mechanics properties (Yang et al. 2019; Liao et al. 2019; Shindo and Matsumoto 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; Withnall et al. 2020; Yuan and Ji 2020; Huang et al. 2020). Most of the recent works on molecular predictions are based on GNN (Duvenaud et al. 2015; Mansimov et al. 2019; Feinberg et al. 2020, 2018). 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 surge in cross-modality contrastive learning (Radford et al. 2021; Wang et al. 2022b; Yang et al. 2022) has spurred its integration into molecular representation. Noteworthy studies such as (Stärk et al. 2022; Zhu et al. 2022) have harnessed contrastive learning to fuse 3D and 2D molecule representations. This technique has found applications in diverse domains, spanning chemical reactions (Lee et al. 2021; Seidl et al. 2022), natural language (Su et al. 2022; Zeng et al. 2022; Edwards, Zhai, and Ji 2021; Seidl et al. 2023), microscopy images (Sanchez-Fernandez et al. 2022), and chemical element knowledge (Fang et al. 2023). Uniquely, our work leverages contrastive learning to facilitate knowledge transfer between biomedical KGs and molecules.

Fusing KG and Molecules. Regarding the amalgamation of KG and molecules, Ye et al. (2021) introduced an approach that blends the static KG embedding of drugs with their structural representations for downstream tasks. However, this method overlooks contextual cues around molecule nodes, thus yielding limited performance enhancements. In a different vein, Wang et al. (2022a) proposed a Graph-of-Graph technique, augmenting graph representation to potentially enrich molecular graph information. Yet, strategies such as pre-training and contrastive learning for aligning the same entity across diverse graph modalities remain unexplored. In contrast, Fang et al. (2022, 2023) pioneered a contrastive learning-based approach, augmenting molecule structures with element-wise knowledge to create an innovative graph structure. This avenue yielded notable advances in molecule property predictions.

Unlike existing methods, GODE extracts a molecule's sub-graph from the biomedical KG, offering a new representation that links molecular data and KGs.

Method

Before presenting our GODE framework, 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, where entities (or nodes) are connected by relations (or edges).*

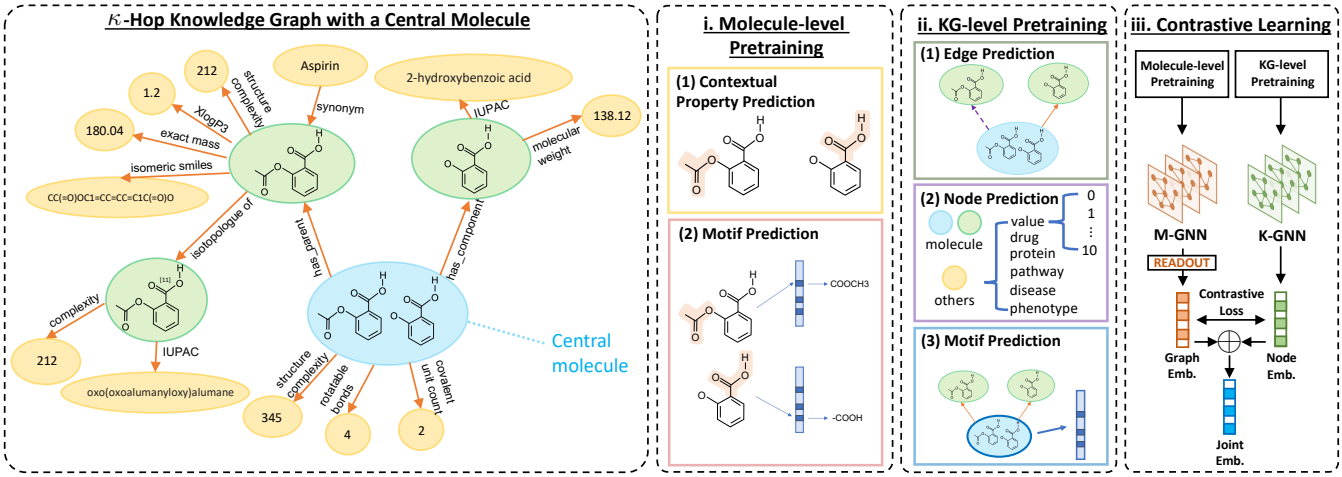

Figure 1: **Overview of our contrastive self-supervised pre-training framework GODE for enhanced molecular representation learning**. ***Left***: the κ-hop KG sub-graph consisting of molecule-relevant relational knowledge, originated by a central molecule. ***Right***: We conduct **(i) Molecule-level Pre-training** on the molecule graphs with masking 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; and **(iii) Contrastive Learning** maximizing the agreement between M-GNN and K-GNN, pre-trained by (i) and (ii), respectively. The joint embedding concatenates the M-GNN graph and K-GNN node embeddings, embodying both structural and diverse molecular information for robust molecule representations.

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 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 $f : \mathcal{M} \rightarrow \mathbb{R}^d$ that is capable of encoding a molecule graph (MG) to a vector $\mathbf{h}_{\text{MG}}$ by mean pooling.*

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

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 downstream molecular property prediction tasks. We detail our approach in the subsequent sections, breaking it down step by step.

Bi-level Self-supervised Pre-training

We propose a bi-level self-supervised pre-training framework to pre-train two GNNs: a molecule-level M-GNN and a KG-level K-GNN trained on molecular graphs and molecule-centered KG sub-graphs, respectively.

i. 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 entire molecule graph embedding, represented as $\mathbf{h}_{\text{MG}}$, is also input into an output layer. This layer predicts the presence or absence of functional group motifs. The embedding $\mathbf{h}_{\text{MG}}$ is derived by applying mean pooling to all nodes: $\mathbf{h}_{\text{MG}} = \text{MEAN}(\mathbf{h}_{v_1}, \mathbf{h}_{v_2}, ..., \mathbf{h}_{v_k} | v_1, v_2, ..., v_k \in \mathcal{V}_m)$, where $\mathbf{h}_{v_1}$, $\mathbf{h}_{v_2}$, ..., $\mathbf{h}_{v_k}$ 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 forecasts a binary label vector, indicating the presence or absence of each functional group motif in G_m, which is detected by RDKit (Landrum et al. 2013).

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 the contextual properties of nodes and the functional group motifs' presence or absence in the molecule graph.

$$\mathcal{L}_{\text{M}} = \sum_{v}^{\mathcal{V}'_m} \log P(p_v|\mathbf{h}_v) + \sum_{j=1}^{n} y_j \log P(M_j|\mathbf{h}_{\text{MG}}) + (1 - y_j) \log(1 - P(M_j|\mathbf{h}_{\text{MG}})), \quad (1)$$

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 $\mathbf{h}_{\text{MG}}$ given its molecule graph.

ii. 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 biomedical 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 stop the expansion of a graph branch when 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:

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

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 $G_{\text{sub}(m,\kappa)}$.

We set three tasks for the KG-level pre-training as shown in module ii of Figure 1:

(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.

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

$$\begin{aligned}\mathcal{L}_{\text{K}} = -\Bigg[& \lambda_{\text{edge}} \underbrace{\sum_{(u,v)}^{\mathcal{E}_{\text{sub}(m,\kappa)}} \log P((u,v)'|\mathbf{h}_u \oplus \mathbf{h}_v)}_{\text{edge prediction}} \\ & + \lambda_{\text{mot}} \underbrace{\sum_{j=1}^{n} [y_j \log P(M_j|\mathbf{h}_c) + (1-y_j)\log(1-P(M_j|\mathbf{h}_c))]}_{\text{motif prediction}} \\ & + \lambda_{\text{node}} \underbrace{\sum_{v}^{\mathcal{V}_{\text{sub}(m,\kappa)}} [\log P(v'|\mathbf{h}_v)}_{\text{node prediction}} \Bigg], \end{aligned} \tag{3}$$

where the first term $(u,v)'$ 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, $\log P(M_j|\mathbf{h}_c)$ is the predicted probability of central molecule c has the j-th functional group motif M_j given its embedding $\mathbf{h}_c$. λ_{edge}, λ_{mot}, and λ_{mol} are hyperparameters balancing the importance of different tasks.

After the KG-level pre-training, K-GNN is able to encode a molecule to a vector $\mathbf{h}_{\text{KG}}$ given its sub-graph in the KG.

iii. 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 negative samples (1) by random sampling, and (2) selected from the neighbors of the positive sample in the κ-hop sub-graph, respectively. The loss is defined as follows:

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

where $\text{sim}(f(m_i), g(s_i))) = \frac{\exp(\tau^{-1}\mathbf{h}^{\text{T}}_{\text{MG}(i)}\mathbf{h}_{\text{KG}(i)})}{\exp(\tau^{-1}\mathbf{h}^{\text{T}}_{\text{MG}(i)}\mathbf{h}_{\text{KG}(i)})+1}$, 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 a hyperparameter indicating the inverse temperature.

Fine-tuning for Downstream Tasks

Upon completing molecule- and KG-level pre-training combined with contrastive learning, we obtain two GNN en-

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

# Triples: 2523867 **# Entities**: 184819 **# Relations**: 39 **# Entity Types**: 7 **# Molecules**: 65454	
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, cooccurence_molecule_molecule, cooccurence_molecule_disease, cooccurence_molecule_gene/protein, neighbor_2d, neighbor_3d, has_same_connectivity, has_component, has_isotopologue, has_parent, has_stereoisomer, to_drug, closematch, type, in_pathway	

coders, f and g, which respectively encode molecules and KG sub-graphs into vectors. For downstream tasks, these encoders are utilized. Specifically, a joint representation is formed by concatenating the embeddings of a molecule graph and its associated KG sub-graph: $\mathbf{h}_{\text{joint}} = \mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$, with $\oplus$ representing concatenation. If the knowledge graph is absent or provides sparse molecule information, only $\mathbf{h}_{\text{MG}}$ is used for prediction. This representation is then utilized to predict the target property y using a multi-layer perception (MLP) with an appropriate activation function. For multi-label classification, we employ binary cross-entropy loss with sigmoid activation, and for regression, we use Mean Squared Error (MSE) loss.

Experiments

Experimental Setting

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 partition this dataset into two subsets with a 9:1 ratio, for training and validation, respectively.

Knowledge Graph-Level Pre-training Data. For the KG-level GNN (K-GNN), we select knowledge graph 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 in Table 1. The dataset is divided into training and validation sets with a 9:1 ratio. The detailed construction of the dataset is outlined in the appendix.

Bi-level Contrastive Learning. To harmonize molecule embeddings from M-GNN and KG-GNN, we set the negative/positive sample ratio as $\alpha = \frac{|\mathcal{D}^-|}{|\mathcal{D}^+|} = 32$ 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.

Downstream Tasks and 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 methodologies employed in previous studies (Rong et al. 2020; Fang et al. 2023).

Implementation. We implement GROVER (Rong et al. 2020) as the M-GNN for molecule-level pre-training and GINE (Hu et al. 2019) for KG-level pre-training. KG embeddings are initialized using TransE over 10 learning epochs. We set $\lambda_{\text{edge}} = 1.5$, $\lambda_{\text{mot}} = 1.8$, and $\lambda_{\text{node}} = 1.5$ shown in Eq. (3). Early stopping is performed based on validation loss. We set the hidden size to 1,200 for both M-GNN and K-GNN. We set the temperature $\tau = 1.0$. We use Adam (Kingma and Ba 2014) optimizer with the Noam learning rate scheduler (Vaswani et al. 2017). All tests are performed on a setup featuring two AMD EPYC 7513 32-Core Processors, 528GB RAM, 8 NVIDIA RTX A6000 GPUs, and CUDA 11.7.

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

Performance of Molecule Property Prediction

Tables 2 and 4 show comparative analyses of performance metrics for classification and regression tasks, respectively. Evidently, our proposed approach, GODE, consistently excels beyond the baseline models across most tasks. In classification tasks, KANO emerges as the only model demonstrating a performance competitive to our method consistently. As a knowledge-driven model, KANO enhances molecular structure by integrating chemical elements' knowledge from its ElementKG. This finding emphasizes the significant benefit of incorporating external knowledge in molecular property prediction. Regarding regression tasks, our model also resides within the top three performers and achieves SOTA results on two of those. Regardless of the task's nature, our model's high-performance

[1]https://moleculenet.org/datasets-1

Table 2: ROC-AUC performance on six *classification* benchmarks (*higher is better*). We report the mean and standard deviation. Top-3 and top-1 results are highlighted in **bold** and **red**, respectively. The backbone M-GNN of GODE is shaded.

Dataset	**BBBP**	**SIDER**	**ClinTox**	**BACE**	**Tox21**	**ToxCast**
# Molecules	2039	1427	1478	1513	7831	8575
# Tasks	1	27	2	1	12	617
GCN (Kipf and Welling 2016)	71.8 ± 0.9	53.6 ± 0.3	62.5 ± 2.8	71.6 ± 2.0	70.9 ± 0.3	65.0 ± 6.1
GIN (Xu et al. 2018)	65.8 ± 4.5	57.3 ± 1.6	58.0 ± 4.4	70.1 ± 5.4	74.0 ± 0.8	66.7 ± 1.5
Weave (Kearnes et al. 2016)	83.7 ± 6.5	54.3 ± 3.4	82.3 ± 2.3	79.1 ± 0.8	74.1 ± 4.4	67.8 ± 2.4
SchNet (Schütt et al. 2017)	84.8 ± 2.2	54.5 ± 3.8	71.7 ± 4.2	76.6 ± 1.1	76.6 ± 2.5	67.9 ± 2.1
MPNN (Gilmer et al. 2017)	91.3 ± 4.1	59.5 ± 3.0	87.9 ± 5.4	81.5 ± 4.4	80.8 ± 2.4	69.1 ± 1.3
DMPNN (Yang et al. 2019)	$\mathbf{91.9 \pm 3.0}$	63.2 ± 2.3	89.7 ± 4.0	85.2 ± 5.3	$\mathbf{82.6 \pm 2.3}$	$\mathbf{71.8 \pm 1.1}$
MGCN (Lu et al. 2019)	85.0 ± 6.4	55.2 ± 1.8	63.4 ± 4.2	73.4 ± 3.0	70.7 ± 1.6	66.3 ± 0.9
MGSSL (Zhang et al. 2021)	70.5 ± 1.1	$\mathbf{64.1 \pm 0.7}$	80.7 ± 2.1	79.7 ± 0.8	76.4 ± 0.4	64.1 ± 0.7
N-GRAM (Liu, Demirel, and Liang 2019)	91.2 ± 1.3	63.2 ± 0.5	85.5 ± 3.7	$\mathbf{87.6 \pm 3.5}$	76.9 ± 2.7	-
HU. et.al (Hu et al. 2019)	70.8 ± 1.5	62.7 ± 0.8	72.6 ± 1.5	84.5 ± 0.7	78.7 ± 0.4	65.7 ± 0.6
$\text{GROVER}_{\text{Large}}$ (Rong et al. 2020) (our M-GNN)	89.0 ± 1.1	56.0 ± 1.8	73.7 ± 4.3	80.1 ± 2.6	73.9 ± 0.7	66.6 ± 2.4
MolCLR (Wang et al. 2021b)	73.3 ± 1.0	61.2 ± 3.6	$\mathbf{89.8 \pm 2.7}$	82.8 ± 0.7	74.1 ± 5.3	65.9 ± 2.1
KANO (Fang et al. 2023)	$\mathbf{92.6 \pm 2.3}$	$\mathbf{63.8 \pm 1.2}$	$\mathbf{91.6 \pm 0.6}$	$\mathbf{90.4 \pm 1.5}$	$\mathbf{81.2 \pm 1.8}$	$\mathbf{72.1 \pm 1.8}$
GODE (ours)	$\mathbf{93.4 \pm 1.0}$	$\mathbf{64.9 \pm 1.3}$	$\mathbf{91.5 \pm 1.4}$	$\mathbf{85.8 \pm 1.6}$	$\mathbf{83.8 \pm 1.9}$	$\mathbf{69.2 \pm 1.1}$

Table 3: **Ablation study of GODE on classification tasks.** Top-3 and top-1 results are highlighted in **bold** and **red**, respectively. The best setting is shaded.

Variants						**BBBP**	**SIDER**	**ClinTox**	**BACE**	**Tox21**	**ToxCast**
Case	**KGE**	**κ-hop**	**Pret.**	**Cont.**	**Embedding**						
①	✓	✗	✗	✗	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KGE}}$	89.2 ± 2.2	62.7 ± 2.1	89.5 ± 3.5	81.7 ± 1.6	77.4 ± 2.9	62.4 ± 1.7
②	✗	$\kappa = 2$	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$	92.6 ± 1.5	$\mathbf{64.8 \pm 0.6}$	90.2 ± 1.9	$\mathbf{84.3 \pm 1.7}$	77.8 ± 2.3	66.8 ± 2.3
③	✓	$\kappa = 2$	✓	✗	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$	$\mathbf{93.5 \pm 1.3}$	63.1 ± 1.5	$\mathbf{90.6 \pm 2.1}$	83.7 ± 2.5	$\mathbf{81.4 \pm 2.3}$	68.0 ± 1.4
④	✓	$\kappa = 2$	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$	$\mathbf{93.4 \pm 1.0}$	$\mathbf{64.9 \pm 1.3}$	$\mathbf{91.5 \pm 1.4}$	$\mathbf{85.8 \pm 1.6}$	$\mathbf{83.8 \pm 1.9}$	$\mathbf{69.2 \pm 1.1}$
⑤	✓	$\kappa = 2$	✓	✓	$\mathbf{h}_{\text{MG}}$	92.9 ± 0.9	$\mathbf{63.3 \pm 1.1}$	$\mathbf{91.5 \pm 2.2}$	83.2 ± 1.1	78.7 ± 2.0	$\mathbf{69.6 \pm 1.9}$
⑥	✓	$\kappa = 3$	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$	$\mathbf{93.1 \pm 1.1}$	61.5 ± 2.3	90.1 ± 1.2	$\mathbf{85.9 \pm 1.8}$	$\mathbf{80.3 \pm 1.9}$	68.8 ± 4.1
⑦	✓	$\kappa = 3$	✓	✓	$\mathbf{h}_{\text{MG}}$	92.5 ± 1.9	62.8 ± 0.7	90.3 ± 1.9	82.4 ± 3.1	77.9 ± 0.6	$\mathbf{69.0 \pm 2.7}$

levels demonstrate its versatility and robustness. The overall relative improvement is 16.1% on all tasks (10.3% for classification and 23.2% for regression tasks). To analyze the effects of variants in GODE, we conduct ablation studies in Table 3 and 5, which are discussed in detail as follows.

Effect of Integration of MolKG To examine the effectiveness of our constructed biomedical KG - MolKG on property prediction tasks, we compare Case ① with our backbone M-GNN GROVER. In Case ①, the molecule embedding, pre-trained by M-GNN (GTransformer in our setting), is combined with the KG embedding trained via the KGE method. It is observed that the inclusion of KG embedding enhances performance on most tasks. However, performance on certain tasks (e.g., ToxCast, ESOL) drops, indicating that the knowledge derived from biomedical KG is not fully acknowledged with such implementation.

KG-level Pre-training and Contrastive Learning The comparative analysis between Case ① and ③ elucidates the impact of KG-level pre-training. Case ③ supplements ① with additional K-GNN pre-training. The analysis unambiguously reveals that K-GNN pre-training substantially bolsters the model's performance. This results in 10.0% overall relative improvements (3.9% for classification and 17.3% in regression tasks), which underscore the effectiveness and potential benefits of incorporating KG-level pre-training in molecular property prediction tasks. Moreover, a comparison between Case ③ and ④ showcases that the introduction of contrastive learning can further optimize performance across all tasks. This is indicated by 9.2% overall relative improvements (1.8% for classification and 18.1% for regression tasks). This evidence suggests that employing contrastive learning to enhance M-GNN and K-GNN embeddings can facilitate more precise predictions, thereby improving the model's overall effectiveness.

Embedding Initialization and Hidden Dimension As per the analysis depicted in Figure 2, we discern that KGE embeddings play a crucial role in enhancing the performance of K-GNN pre-training tasks (Eq. (3)). This enhancement is evident in the reduced validation loss throughout the training process, signifying more accurate predictions. Further, the data suggests a positive correlation between the dimensionality of the embeddings and the quality of pre-training results, where higher dimensions tend to yield better outcomes. In addition, Table 3 and 5 reveal that incorporating KGE embeddings (Case ④) generally leads to superior performance, always outperforming the cases where KGE embeddings are omitted (Case ②). These findings underscore the vital role of embedding initialization via KGE methods in optimizing performance during molecular property prediction tasks.

Efficacy of Knowledge Transfer We evaluated the effectiveness of applying contrastive learning for domain knowl-

Table 4: RMSE (for FreeSolv, ESOL, Lipophilicity) and MAE (for QM7/8) performance on five *regression* benchmarks (*lower is better*). Top-3 and top-1 results are highlighted in **bold** and **red**, respectively. The backbone M-GNN of GODE is shaded.

Datasets	**FreeSolv**	**ESOL**	**Lipophilicity**	**QM7**	**QM8**
# Molecules	642	1128	4200	6830	21786
# Tasks	1	1	1	1	12
GCN (Kipf and Welling 2016)	2.870 ± 0.140	1.430 ± 0.050	$\mathbf{0.712 \pm 0.049}$	122.9 ± 2.2	0.037 ± 0.001
GIN (Xu et al. 2018)	2.765 ± 0.180	1.452 ± 0.020	0.850 ± 0.071	124.8 ± 0.7	0.037 ± 0.001
Weave (Kearnes et al. 2016)	2.398 ± 0.250	1.158 ± 0.055	0.813 ± 0.042	94.7 ± 2.7	0.022 ± 0.001
SchNet (Schütt et al. 2017)	3.215 ± 0.755	1.045 ± 0.064	0.909 ± 0.098	$\mathbf{74.2 \pm 6.0}$	0.020 ± 0.002
MPNN (Gilmer et al. 2017)	$\mathbf{1.621 \pm 0.952}$	1.167 ± 0.430	$\mathbf{0.672 \pm 0.051}$	111.4 ± 0.9	$\mathbf{0.015 \pm 0.001}$
DMPNN (Yang et al. 2019)	1.673 ± 0.082	1.050 ± 0.008	$\mathbf{0.683 \pm 0.016}$	103.5 ± 8.6	$\mathbf{0.016 \pm 0.001}$
MGCN (Lu et al. 2019)	3.349 ± 0.097	1.266 ± 0.147	1.113 ± 0.041	77.6 ± 4.7	0.022 ± 0.002
N-GRAM (Liu, Demirel, and Liang 2019)	2.512 ± 0.190	1.100 ± 0.160	0.876 ± 0.033	125.6 ± 1.5	0.032 ± 0.003
HU. et.al (Hu et al. 2019)	2.764 ± 0.002	1.100 ± 0.006	0.739 ± 0.003	113.2 ± 0.6	0.022 ± 0.001
GROVER (Rong et al. 2020) (our M-GNN)	2.445 ± 0.761	$\mathbf{1.028 \pm 0.145}$	0.843 ± 0.122	133.8 ± 3.8	0.036 ± 0.005
MolCLR (Wang et al. 2021b)	2.301 ± 0.247	1.113 ± 0.023	0.789 ± 0.009	90.0 ± 1.7	0.019 ± 0.013
KANO (Fang et al. 2023)	$\mathbf{1.443 \pm 0.315}$	$\mathbf{0.914 \pm 0.092}$	0.726 ± 0.012	$\mathbf{73.5 \pm 3.5}$	$\mathbf{0.021 \pm 0.002}$
GODE (ours)	$\mathbf{1.495 \pm 0.343}$	$\mathbf{0.910 \pm 0.028}$	0.723 ± 0.102	$\mathbf{73.0 \pm 3.8}$	$\mathbf{0.021 \pm 0.001}$

Table 5: **Ablation study of GODE on regression tasks.** Top-3 and top-1 results are highlighted in **bold** and **red**, respectively. The best setting is shaded.

	Variants					**FreeSolv**	**ESOL**	**Lipophilicity**	**QM7**	**QM8**
Case	**KGE**	**κ-hop**	**Pret.**	**Cont.**	**Embedding**					
①	✓	✗	✗	✗	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KGE}}$	2.341 ± 0.382	1.205 ± 0.114	0.829 ± 0.049	119.9 ± 2.1	0.032 ± 0.003
②	✗	$\kappa = 2$	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$	$\mathbf{1.664 \pm 0.420}$	$\mathbf{0.912 \pm 0.061}$	$\mathbf{0.731 \pm 0.086}$	102.9 ± 1.9	$\mathbf{0.025 \pm 0.003}$
③	✓	$\kappa = 2$	✓	✗	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$	1.995 ± 0.223	$\mathbf{0.998 \pm 0.021}$	0.755 ± 0.093	$\mathbf{80.1 \pm 1.2}$	0.028 ± 0.002
④	✓	$\kappa = 2$	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$	$\mathbf{1.495 \pm 0.343}$	$\mathbf{0.910 \pm 0.028}$	$\mathbf{0.723 \pm 0.052}$	$\mathbf{73.0 \pm 3.8}$	$\mathbf{0.021 \pm 0.001}$
⑤	✓	$\kappa = 2$	✓	✓	$\mathbf{h}_{\text{MG}}$	1.877 ± 0.228	1.012 ± 0.085	$\mathbf{0.723 \pm 0.069}$	92.3 ± 2.9	0.028 ± 0.002
⑥	✓	$\kappa = 3$	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$	$\mathbf{1.744 \pm 0.415}$	1.134 ± 0.112	0.770 ± 0.061	$\mathbf{77.0 \pm 4.9}$	$\mathbf{0.024 \pm 0.004}$
⑦	✓	$\kappa = 3$	✓	✓	$\mathbf{h}_{\text{MG}}$	1.779 ± 0.191	1.007 ± 0.226	$\mathbf{0.758 \pm 0.029}$	85.4 ± 3.3	0.030 ± 0.001

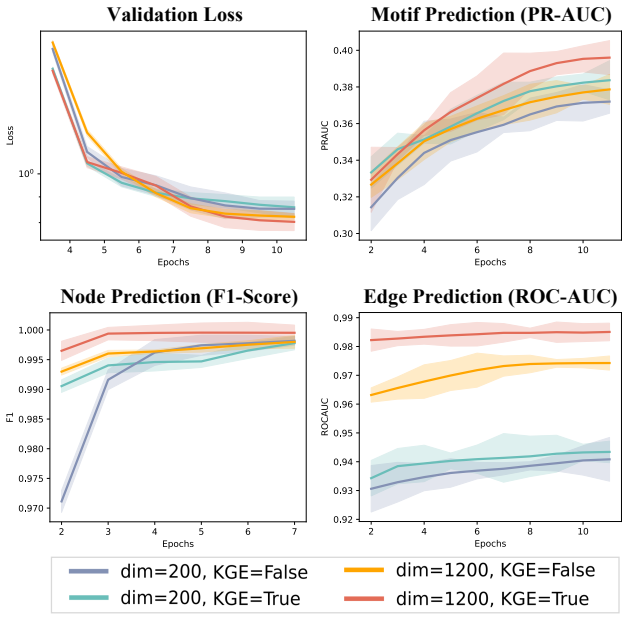

Figure 2: **Performance of K-GNN pre-training tasks.** We report the means and standard deviation based on five runs with different random seeds.

edge transfer from the biomedical KG to the molecular representation ($\mathbf{h}_{\text{MG}}$). This was done through a comparative analysis of GROVER and Cases ④, ⑤, ⑥, and ⑦. The study uncovers that the performance of the embedding derived solely from GODE's M-GNN may not reach the same level as that of the bi-level concatenated embedding, yet it significantly surpasses that of the raw M-GNN (GROVER). This enhancement is vividly evidenced by an overall relative improvement of 12.8% (8.2% for classification and 18.3% for regression) by comparing ⑤ and GROVER; and an overall relative improvement of 12.4% (7.4% for classification and 18.4% for regression) by comparing ⑦ and GROVER.

Case Study Within the BBBP dataset, the molecule acetylsalicylate (commonly referred to as aspirin) could not be accurately predicted by our M-GNN model or by the methods in Case ① or ③. Yet, in Case ④, harnessing relational knowledge from its KG sub-graph (e.g., [*acetylsalicylate, may treat, neurological conditions*], [*acetylsalicylate, is, lipophilic*]) and employing contrastive learning, accurate predictions were achieved. This instance underscores the efficacy of our bi-level self-supervised pre-training with contrastive learning in nuanced molecular property predictions.

Conclusion

In this paper, we have developed a molecule-knowledge graph bi-level self-supervised pre-training and contrastive learning framework - GODE, to enhance the molecule representation with biomedical domain knowledge, for molecular property prediction tasks. We have conducted thorough empirical studies to validate the effectiveness of our proposed method and its variants.

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.

Belleau, F.; Nolin, M.-A.; Tourigny, N.; Rigault, P.; and Morissette, J. 2008. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41(5): 706–716.

Blum, L. C.; and Reymond, J.-L. 2009. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society*, 131(25): 8732–8733.

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.

Cai, H.; Zhang, H.; Zhao, D.; Wu, J.; and Wang, L. 2022. FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction. *Briefings in Bioinformatics*, 23(6): bbac408.

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.

Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; and Jensen, K. F. 2019. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, 10(2): 370–377.

Delaney, J. S. 2004. ESOL: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3): 1000–1005.

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.

Edwards, C.; Zhai, C.; and Ji, H. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 595–607.

Fang, Y.; Zhang, Q.; Yang, H.; Zhuang, X.; Deng, S.; Zhang, W.; Qin, M.; Chen, Z.; Fan, X.; and Chen, H. 2022. Molecular Contrastive Learning with Chemical Element Knowledge Graph. arXiv:2112.00544.

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.

Feinberg, E. N.; Joshi, E.; Pande, V. S.; and Cheng, A. C. 2020. Improvement in ADMET prediction with multitask deep featurization. *Journal of medicinal chemistry*, 63(16): 8835–8848.

Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; and Pande, V. S. 2018. PotentialNet for molecular property prediction. *ACS central science*, 4(11): 1520–1530.

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.

Gayvert, K. M.; Madhukar, N. S.; and Elemento, O. 2016. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology*, 23(10): 1294–1301.

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.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 1026–1034.

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.; Fu, T.; Glass, L. M.; Zitnik, M.; Xiao, C.; and Sun, J. 2020. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23): 5545–5547.

Huang, K.; and Zitnik, M. 2020. Graph meta learning via local subgraphs. *Advances in neural information processing systems*, 33: 5862–5874.

Huang, R.; and Xia, M. 2017. Editorial: Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways As Mediated by Exposure to Environmental Toxicants and Drugs. *Frontiers in Environmental Science*, 5.

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*.

Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; and Riley, P. 2016. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30: 595–608.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kuhn, M.; Letunic, I.; Jensen, L. J.; and Bork, P. 2016. The SIDER database of drugs and side effects. *Nucleic acids research*, 44(D1): D1075–D1079.

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.

Martins, I. F.; Teixeira, A. L.; Pinheiro, L.; and Falcao, A. O. 2012. A Bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, 52(6): 1686–1697.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mobley, D. L.; and Guthrie, J. P. 2014. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28: 711–720.

Piñero, J.; Ramírez-Anguita, J. M.; Saüch-Pitarch, J.; Ronzano, F.; Centeno, E.; Sanz, F.; and Furlong, L. I. 2020. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1): D845–D855.

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.

Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; and Von Lilienfeld, O. A. 2015. Electronic spectra from TDDFT and machine learning in chemical space. *The Journal of chemical physics*, 143(8).

Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; et al. 2016. ToxCast chemical landscape: paving the road to 21st century toxicology. *Chemical research in toxicology*, 29(8): 1225–1251.

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*.

Shindo, H.; and Matsumoto, Y. 2019. Gated graph recursive neural networks for molecular property prediction. *arXiv preprint arXiv:1909.00259*.

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*.

Subramanian, G.; Ramsundar, B.; Pande, V.; and Denny, R. A. 2016. Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling*, 56(10): 1936–1949.

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*.

Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; et al. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1): D607–D613.

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. 2022a. Imbalanced graph classification via graph-of-graph neural networks. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2067–2076.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022b. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Withnall, M.; Lindelöf, E.; Engkvist, O.; and Chen, H. 2020. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. *Journal of cheminformatics*, 12(1): 1–18.

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, 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, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; and Huang, J. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15671–15680.

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*.

Zeng, Z.; Yao, Y.; Liu, Z.; and Sun, M. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1): 862.

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.

Zhu, J.; Xia, Y.; Wu, L.; Xie, S.; Qin, T.; Zhou, W.; Li, H.; and Liu, T.-Y. 2022. Unified 2d and 3d pre-training of molecular representations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2626–2636.

Knowledge Graph Construction and Process

The construction of our molecule-centric knowledge graph - MolKG, involved a comprehensive data retrieval process of knowledge graph triples relevant to molecules. We retrieve the data from two distinguished sources: PubChemRDF[2] (Fu et al. 2015) and PrimeKG (Chandak, Huang, and Zitnik 2023a). From PubChemRDF, we concentrated on triples from six specific subdomains:

- *Compound*: This encompasses compound-specific relation types such as *parent compound*, *component compound*, and *compound identity group*.
- *Cooccurrence*: This domain captures triples like *compound-compound*, *compound-disease*, and *compound -gene* co-occurrences. By ranking co-occurrences based on their scores, we selected the top 5 compounds, diseases, and genes for each molecule, resulting in at most 15 co-occurred entities per molecule.
- *Descriptor*: This domain details explicit molecular properties including *structure complexity*, *rotatable bond*, and *covalent unit count*.
- *Neighbors*: Represents the top N molecules similar in 2-d and 3-d structures. For our dataset, we integrated the top 3 similar molecules from both 2-d and 3-d structures for each molecule.
- *Component*: Associates molecules with their constituent components.
- *Same Connectivity*: Showcases molecules with identical connectivity to source molecules.

From PrimeKG, we pursued a rigorous extraction technique, deriving 3-hop sub-graphs for all 7,957 drugs, regarded as molecules, from the entirety of the knowledge graph. Consistency and accuracy in data handling were paramount. We utilized recognized information retrieval tools[3][4] to bridge various representations and coding paradigms for identical molecular entities. Compound ID (CID) served as our go-to medium for molecular conversions across the two knowledge graphs. Lastly, within our assembled knowledge graph, entities identified as 'value' underwent a normalization process. Subsequently, we classified these entities, ensuring a maximum class count of 10.

Datasets of Downstream Tasks

We delve deeper into the various molecular property prediction datasets:

- **BBBP** (Martins et al. 2012): The Blood-Brain Barrier Penetration (BBBP) dataset is crucial for drug discovery, especially for neurological disorders. It characterizes whether a chemical compound can cross the blood-brain barrier, which is essential for a compound's efficacy in treating brain disorders. This dataset aids researchers in predicting and understanding the molecular properties that influence barrier penetration.
- **SIDER** (Kuhn et al. 2016): The Side Effect Resource (SIDER) provides information about marketed medications and their observed adverse effects. Such insights are essential for pharmacovigilance, enabling researchers to discern patterns or predict potential side effects of new compounds based on their molecular structures or mechanisms of action.
- **ClinTox** (Gayvert, Madhukar, and Elemento 2016): This dataset serves as a comparative tool, distinguishing between drugs that gained FDA approval and those that faced rejections due to toxicological concerns during clinical evaluations. By analyzing these disparities, researchers can better understand and anticipate toxicological profiles of novel compounds.
- **BACE** (Subramanian et al. 2016): BACE dataset offers insights into compounds and their potential as inhibitors for human β-secretase 1 (BACE-1), an enzyme linked to Alzheimer's disease. Compounds that inhibit BACE-1 could play a role in developing treatments for Alzheimer's, making this dataset pivotal for neurological drug discovery.
- **Tox21** (Huang and Xia 2017): This publicly accessible resource gives a comprehensive look into the toxicity profiles of various compounds. It played a central role in the 2014 Tox21 Data Challenge, where scientists aimed to develop models to predict toxic responses more effectively, underscoring the need for safer drug design.
- **ToxCast** (Richard et al. 2016): With an array of toxicity labels from high-throughput screenings, ToxCast offers a rich resource for understanding the toxicological profiles of thousands of compounds. Such broad-scale screenings enable swift evaluations, guiding researchers in the early stages of drug development.
- **FreeSolv** (Mobley and Guthrie 2014): A dataset that brings together information on the hydration free energy of molecules in water. The dual presence of experimental data and alchemical free energy calculations offers researchers a robust platform to understand solvation processes and predict such properties for novel molecules.
- **ESOL** (Delaney 2004): Understanding the solubility of compounds is fundamental in drug formulation and delivery. The ESOL dataset chronicles solubility attributes, providing a structured framework to predict and modify solubility properties in drug design.
- **Lipophilicity** (Gaulton et al. 2012): Extracted from the ChEMBL database, this dataset focuses on a compound's affinity for lipid bilayers—a key factor in drug absorption and permeability. It provides valuable insights derived from octanol/water distribution coefficient experiments.
- **QM7** (Blum and Reymond 2009): A curated subset of GDB-13, the QM7 dataset houses details on computed atomization energies of stable, potentially synthesizable organic molecules. It provides an arena for validating quantum mechanical methods against empirical data, bridging computational studies with experimental chemistry.

[2]https://pubchem.ncbi.nlm.nih.gov/docs/rdf-intro

[3]https://pubchem.ncbi.nlm.nih.gov/docs/pug-rest

[4]https://www.ncbi.nlm.nih.gov/home/develop/api/

- **QM8** (Ramakrishnan et al. 2015): A more extensive dataset, QM8 encompasses computer-generated quantum mechanical properties. It details aspects like electronic spectra and the excited state energy of molecules, offering a robust resource for computational chemists aiming to predict or understand such attributes.

Experimental Details

For the negative sampling of contrastive learning, we sample $\alpha = \frac{|\mathcal{D}^-|}{|\mathcal{D}^+|}$ from $\{2, 4, 8, 16, 32, 64\}$. The balancing parameters λ_{edge}, λ_{mot}, and λ_{node} for K-GNN pre-training are adjusted within the range of 1.0 to 2.0, using increments of 0.1. We choose the hidden dimension of the K-GNN embedding from $\{200, 400, 800, 1200\}$. Following previous works (Rong et al. 2020; Fang et al. 2023), we use RDKit to extract additional features of M-GNN. The hidden dimension of M-GNN is tuned in the range of $\{400, 800, 1200\}$. The M-GNN is pre-trained for 500 epochs with a learning rate of 1.5×10^{-4} and weight decay of 10^{-7}. PReLU (He et al. 2015) is used as the activation function for M-GNN. When M-GNN is GROVER, we take both "atom from atom" and "atom from bond" as the molecule representation.

Justifications for GODE

In the proposed methodology, we aim to construct a powerful molecule representation via a bi-level self-supervised pre-training technique that leverages both molecular graphs (M-GNN) and Knowledge Graphs (K-GNN). To bridge these two representations and leverage the strengths of both, contrastive learning is used. To validate and support the proposed methodologies mathematically, the following are the detailed justifications and explanations:

Justifications for Bi-level Pre-training

Molecule-level Pre-training The objective for molecule-level pre-training is to capture local atom properties (contextual property prediction) and global functional group motifs (graph-level motif prediction), as described by Eq. (1). The goal is to maximize the likelihood of the true contextual property and the motif labels given their embeddings.

The first term, $\log P(p|\mathbf{h}_v)$ in Eq. (1), is a direct log-likelihood of the true contextual property given the node embedding. Maximizing this term encourages the GNN to capture local structural information of atoms in the molecule graph. The second and third terms work in tandem for each possible motif M_j. If the motif M_j is present (i.e., $y_j = 1$), we want to maximize $P(M_j|\mathbf{h}_{\text{MG}})$. If the motif M_j is absent (i.e., $y_j = 0$), we want to maximize $1 - P(M_j|\mathbf{h}_{\text{MG}})$. This is achieved via maximizing the combined term $y_j \log P(M_j|\mathbf{h}_{\text{MG}}) + (1 - y_j) \log(1 - P(M_j|\mathbf{h}_{\text{MG}}))$.

In maximizing this loss, we ensure that our M-GNN captures both the local properties of atoms and the global properties (functional motifs) of the molecule.

KG-level Pre-training. The proposed loss function for the K-GNN (Eq. (3)) encapsulates three main tasks: edge prediction, motif prediction, and node prediction.

The term $\lambda_{\text{edge}} \sum_{(u,v)}^{\mathcal{E}_{\text{sub}(m,\kappa)}} \log P((u,v)'|\mathbf{h}_u \oplus \mathbf{h}_v)$ ensures the GNN captures the relationship between two nodes. Maximizing this log-likelihood encourages the K-GNN to capture semantic meanings and relationships between entities in the KG. The motif prediction task encourages the K-GNN to capture the properties of the central molecule, much like the motif prediction in M-GNN but now in the context of a knowledge graph. By maximizing the likelihood of the motif labels, the GNN captures the molecular motifs in the context of surrounding information from the KG. The node prediction task helps the K-GNN understand the semantic roles of individual nodes in the sub-graph.

By maximizing this combined loss, the K-GNN captures edge semantics, node roles, and molecular motifs in the context of a KG.

Justifications for Contrastive Learning

Contrastive learning inherently aligns representations originating from disparate sources. For our context, this involves the M-GNN and K-GNN systems. By ensuring that representations of the same molecule from both platforms are more proximate in latent space and concurrently distancing representations of distinct molecules, we establish an efficient mechanism for knowledge interchange.

Consider representations $\mathbf{h}_{\text{MG}}$ derived from M-GNN and $\mathbf{h}_{\text{KG}}$ from K-GNN. The similarity metric between these representations for a positively correlated pair is represented by $s(\mathbf{h}_{\text{MG}}, \mathbf{h}_{\text{KG}})$. The overarching objective of the contrastive loss is to optimize:

$$s(\mathbf{h}_{\text{MG}}, \mathbf{h}_{\text{KG}}) - \mathbb{E}_{\text{neg}}[s(\mathbf{h}_{\text{MG}}, \mathbf{h}_{\text{neg}})], \tag{5}$$

where $\mathbb{E}_{\text{neg}}[.]$ stands for the anticipated similarity over negatively correlated pairs. By maximizing this difference, the collective knowledge—such as shared motifs and properties—across both M-GNN and K-GNN becomes intrinsically woven into their respective representations.

By applying the InfoNCE loss as illustrated in Eq. (4), we ensure a refined alignment between the M-GNN and K-GNN representations. This alignment seeks to minimize the InfoNCE loss, guaranteeing that representations of identical molecules from the two models approach one another in latent space, thereby amplifying $s(\mathbf{h}_{\text{MG}}, \mathbf{h}_{\text{KG}})$, while representations of unalike molecules are distanced by reducing their similarity to $\mathbf{h}_{\text{neg}}$.

Enhancement in Fine-tuning

For downstream tasks, when these aligned representations are used, there are two primary advantages:

Richer Information Source. As the molecular representation $\mathbf{h}_{\text{MG}}$ gets influenced by the knowledge graph representation $\mathbf{h}_{\text{KG}}$, it captures not only local structural details but also broader semantic information from the KG. The concept of entropy, which quantifies the amount of uncertainty (or information) associated with a representation, can shed light on this. Consider the joint entropy of the combined representations:

$$H(\mathbf{h}_{\text{MG}}, \mathbf{h}_{\text{KG}}) \tag{6}$$

This joint entropy captures (1) Information exclusive to the molecular graph; (2) information exclusive to the knowledge graph; (3) any overlapping or shared information. Therefore, this combined representation contains richer informa-

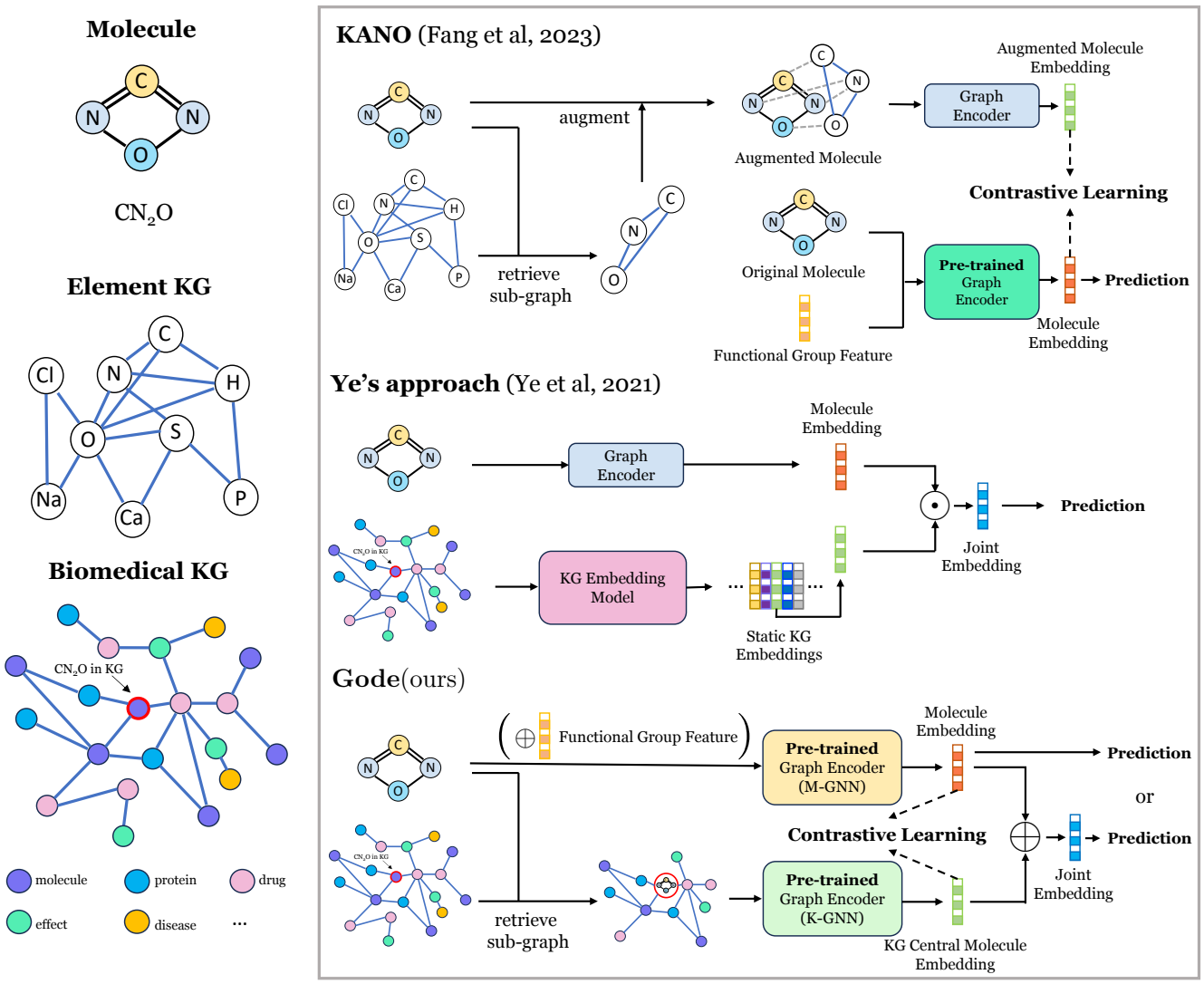

Figure 3: **An overview of the difference between GODE with similar works (Ye et al. 2021; Fang et al. 2023) leveraging both knowledge graph and molecule.** Details such as pre-training strategies or KG embedding initialization are not depicted, for clearer presentations.

tion content than molecular representation alone. Mathematically, this is captured by:

$$H(\mathbf{h}_{\text{MG}}, \mathbf{h}_{\text{KG}}) \geq H(\mathbf{h}_{\text{MG}}) \tag{7}$$

implying that the uncertainty (or richness) of the combined representation is at least as much as that of the molecular representation alone.

Regularization Effect Aligning representations from two sources can have a regularization effect, reducing the risk of overfitting in downstream tasks. This is because the representations are now not only optimized for one source but are balanced to be optimal for both.

We denote our fine-tuning task's loss function as $\mathcal{L}_{\text{task}}$. In a framework influenced by contrastive learning, the representations are harmonized from both sources:

$$\mathbf{h} = f(\mathbf{h}_{\text{MG}}, \mathbf{h}_c), \tag{8}$$

where f is an MLP, that harmonizes the two representations. Inserting this into the loss function:

$$\mathcal{L}_{\text{task}}(f(\mathbf{h}_{\text{MG}}, \mathbf{h}_c)) \tag{9}$$

while the loss would be $\mathcal{L}_{\text{task}}(\mathbf{h}_{\text{MG}})$ for the scenario without the transfer of knowledge. To comprehend the advantage, we consider the gradients for both scenarios:

$$\nabla\mathcal{L}_{\text{task}}(f(\mathbf{h}_{\text{MG}}, \mathbf{h}_{\text{KG}})), \quad \text{and} \quad \nabla\mathcal{L}_{\text{task}}(\mathbf{h}_{\text{MG}}) \tag{10}$$

The former gradient, influenced by both molecular and knowledge graphs, results in a diversified and holistic understanding. Comparatively, the latter is singularly shaped by the molecular graph. As such, the variance of the gradient with combined influence is typically lower:

$$\text{Var}\left(\nabla\mathcal{L}_{\text{task}}(f(\mathbf{h}_{\text{MG}}, \mathbf{h}_{\text{KG}}))\right) \leq \text{Var}\left(\nabla\mathcal{L}_{\text{task}}(\mathbf{h}_{\text{MG}})\right) \tag{11}$$

This reduced variance ensures that updates during training are stable and consistent, fostering improved generalization. Alongside, the smoother loss landscape that emerges from integrating both sources further contributes to enhanced model generalization, as it discourages settling into sharp, non-generalizable minima.

In summary, contrastive learning's ability to amalgamate the strengths of both the molecular graph and knowledge graph results in a richer, more stabilized representation. This not only introduces an inherent regularization but also ensures a holistic absorption of information, thereby enhancing the fine-tuning performance.

Comparison with Similar Studies

In this section, we compare the proposed GODE method with some similar studies integrating knowledge graph and molecule for molecular property predictions. Specifically, we compare to (Ye et al. 2021) and (Fang et al. 2023), as shown in Figure 3.

KANO (Fang et al. 2023) presents Element KG, a knowledge graph detailing the relational connections among chemical elements. In their approach to transfer knowledge from this KG to molecular representations, the process begins by extracting an element sub-graph tailored to a specific molecule. This sub-graph is subsequently integrated with the original molecule graph, effectively enriching the atomic structures within the molecule using the KG. For the encoding phase, they use a non-pre-trained graph encoder to derive the embedding of the enhanced molecule structure. In parallel, they utilize a pre-trained graph encoder to capture the graph embedding of the molecule, with features sourced from RDKit. The culmination of this process is the application of contrastive learning, aligning the embedding of the supplemented molecule with the original molecule's embedding, which is then used to fine-tune downstream tasks. This meticulous procedure ensures an effective knowledge transfer from elemental details to the overall molecular representation.

Ye's approach Presented in (Ye et al. 2021), Ye's method was initially developed for recommendation systems. However, its potential extends to predicting molecular properties, as illustrated in Figure 3. The procedure begins by obtaining embeddings for both the molecule and the biomedical knowledge graph, achieved through a molecule graph encoder and a KG embedding technique, respectively. Subsequently, element-wise multiplication is employed to combine these embeddings for predictive tasks. Notably, a primary drawback of Ye's strategy is its reliance on static, global embeddings. This can sometimes neglect the nuanced, local information pertaining to the targeted entity. Furthermore, there is a conspicuous absence of any mechanism to consolidate the same entity represented in different modalities. This omission creates a disconnect in the knowledge transfer from the biomedical KG to the molecular representation.

GODE (ours) On the other hand, our GODE methodology offers a distinct approach to integrating knowledge graphs and molecular structures for enhanced molecular property predictions. Unlike other methods, GODE directly retrieves a sub-graph tailored to the central molecule from the biomedical knowledge graph (KG). This direct retrieval ensures that the most relevant and contextual information from the KG is harnessed. GODE employs two pre-trained graph encoders, K-GNN and M-GNN, where the former is pre-trained on molecule-centric KG sub-graphs, and the latter is pre-trained on the molecule graph's structural information. An optional enhancement (Rong et al. 2020; Fang et al. 2023) to this process is the inclusion of the functional group feature, which is retrieved by RDKit. A pivotal aspect of the GODE method is the alignment process. Through the application of contrastive learning, the representations of the same molecule, as derived from the two distinct graphs (biomedical KG and molecule graph), are meticulously aligned. This alignment ensures that the embeddings are harmonized and that there is a seamless transfer of knowledge between the two representations. In the subsequent fine-tuning stage, GODE offers flexibility. Users can either employ the concatenated embedding, which is a fusion of the outputs from M-GNN and K-GNN, or opt to use only the embedding from M-GNN. This adaptability ensures that the method can be tailored to best suit specific downstream prediction tasks, optimizing accuracy and efficiency.

Broader Impact

The development of GODE offers a significant advancement in the realm of molecular representation learning. Its broader impacts can be summarized as follows:

Enhanced Drug Discovery By providing a robust knowledge-enhanced representation of molecules, GODE can potentially accelerate drug discovery processes. This could lead to faster identification of potential drug candidates and reduce the time and cost associated with bringing new drugs to the market.

Interdisciplinary Applications The fusion of molecular structures with knowledge graphs can be applied beyond the realm of molecular biology. This approach can be extended to other scientific domains where entities have both intrinsic structures and are part of larger networks.

Potential Ethical Considerations As with any predictive model, there's a need to ensure that the data used is unbiased and representative. Misrepresentations or biases in the knowledge graph or molecular data can lead to skewed predictions, which could have implications in real-world applications, especially in drug development.