DISTILLING PRE-TRAINED KNOWLEDGE IN CHEMICAL REACTIONS FOR MOLECULAR PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

How to effectively represent molecules is a long-standing challenge for molecular property prediction and drug discovery. Recently, accumulative unlabelled molecule data have spurred the rapid development of pre-training methods for molecular representation learning. However, these works mainly focus on devising self-supervised learning tasks and/or introducing 3D geometric information based on molecular structures with little chemical domain knowledge involved. To address this issue, we propose a novel method (MolKD) by Distilling pre-trained Knowledge in chemical reactions to assist Molecular property prediction. Specifically, MolKD first learns effective representations by incorporating *reaction yields* to measure transformation efficiency of the reactant-product pair when pre-training on reactions. Next, MolKD introduces the reaction-to-molecule distillation to transfer cross-modal knowledge between pre-training chemical reaction data and the downstream molecular property prediction tasks. Extensive experiments show that our method can learn effective molecular representations, achieving superior performance compared with state-of-the-art baselines, *e.g.*, 2.8% absolute Hit@1 gain on USPTO in chemical reaction prediction and 1.6% absolute AUC-ROC gain on Tox21 with $1/3$ pre-training data size in molecular property prediction. Further investigations on pre-trained molecular representations indicate that MolKD learns to distinguish chemically meaningful molecular similarities, which enables molecular property prediction with high robustness and interpretability.

1 INTRODUCTION

Effective molecular representations plays an important role in AI-aided drug design and discovery, such as chemical reaction prediction (Lu & Zhang, 2022; Bi et al., 2021), molecular property prediction (Shen et al., 2021; Liu et al., 2022a), and molecule generation (Xu et al., 2022; Luo & Ji, 2021). In computational chemistry, researchers have proposed many methods of conventional molecular representations, such as SMILES (Weininger, 1988), SELFIES (Krenn et al., 2020), and ECFP (Rogers & Hahn, 2010) [1].

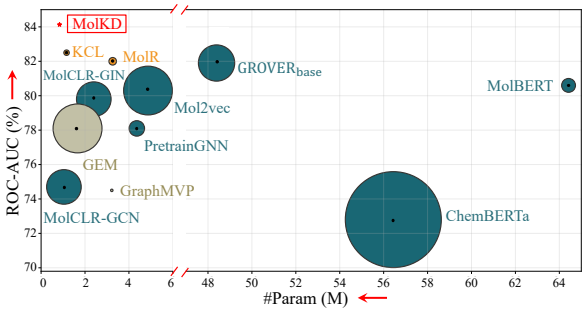

Figure 1: ROC-AUC score *v.s.* the number of model's parameters on Tox21. Each datapoint is visualized as a circle whose radius is proportional to $\sqrt{p}$, where p is the size of pre-training data. The color of the circle denotes the category of pre-training methods as introduced in Sec. 2. We find that MolKD achieves the best performance with a small number of model's parameters and pre-training data.

However, such string-based representations are hard to directly encode the important topology and structural information of a molecule. Because molecules can be naturally represented as graphs by taking atoms as nodes and chemical bonds as edges. In order to preserve rich structural information, many recent works exploit graph neural networks (GNNs) to extract and propagate messages of each atom within

[1] For example, the molecular formula and the SMILES string (Simplified Molecular-Input Line-Entry System) of methyl acetate are C3H6O2 and CC(=O)OC, respectively.

its neighbors, and have shown promising results in molecular property prediction (Vamathevan et al., 2019; Stärk et al., 2022). However, acquiring labeled molecule data usually requires time-consuming laborious and costly wet-lab experiments (Atz et al., 2021; Hao et al., 2020), which hinders the successful application of GNN-based methods in molecule property prediction.

To mitigate the scarcity of labeled molecule data, recent progress has been made by pre-training molecular representations on unlabelled molecule data. Typically, the main idea behind pre-training strategies is to leverage enormous unlabeled data to learn molecular representations fit for the downstream prediction tasks. Wang et al. (2022b); Liu et al. (2022a) design several self-supervised learning tasks to model molecular structures. Fang et al. (2022a); Stärk et al. (2022) further introduce the 3D molecular information and align the 2D graph with the 3D conformation representation of a molecule. More discussions can be found in Sec. 2. However, these methods may suffer from low data efficiency and generalization ability without the help of chemical domain knowledge.

To effectively leverage the benefits of chemical domain knowledge, we propose a novel method (MolKD) by Distilling pre-trained Knowledge in chemical reactions to assist Molecular property prediction. Specifically, *yield* plays a critical role in a chemical reaction by measuring the transformation efficiency of the reactant-product pair. MolKD can incorporate this important factor and learn powerful molecular representations by our proposed yield-guided chemical reaction pre-training method. Moreover, to narrow the gap in data modality between reactions and molecules, we introduce the reaction-to-molecule distillation to transfer cross-modal knowledge between pre-training chemical reaction data and the downstream molecular property prediction tasks.

To validate the high quality of molecular representations, we compare our yield-guided chemical reaction pre-training method with competitive baselines, and achieve 2.8% absolute Hit@1 gain on USPTO in chemical reaction prediction. We also visualize the selected molecules to show that the learned molecular representations are chemically meaningful by encoding structural and synthetic semantics in the representation domain (See Fig 4). To verify the effectiveness of our proposed MolKD, we compare it with several state-of-the-art baselines on 9 molecular property prediction benchmarks, among which MolKD achieves superior performance on 8 challenging tasks, *e.g.*, 1.6% absolute AUC-ROC gain on Tox21 with only $1/3$ pre-training data size (See Fig. 1). We further investigate MolKD on PhysProp (Li et al., 2022a) to demonstrate its robustness and interpretability.

2 Related work

Pre-training for Molecular Representations. In general, the pre-training methods for molecular representations fall into three categories. (1) Wang et al. (2022b); Liu et al. (2022a); Fang et al. (2022b); Hu et al. (2020); Rong et al. (2020); Li et al. (2020) design dedicated self-supervised learning tasks based on string-level (1D) and graph-level (2D) structures of molecule data. Hu et al. (2020) propose the node- (attribute masking) and graph-level (context prediction) pre-training strategies to make accurate and robust predictions on a variety of downstream tasks. GROVER (Rong et al., 2020) adopts the Transformer structure to the designed self-supervised task (*i.e.*, contextual property prediction and graph-level motif prediction). (2) Fang et al. (2022a); Stärk et al. (2022); Liu et al. (2022b); Li et al. (2022a) further introduce 3D conformations of a molecule and align its 2D and 3D representation. GEM (Fang et al., 2022a) incorporates 3D geometric information into several dedicated geometry-level self-supervised learning strategies. GraphMVP (Liu et al., 2022b) introduces a multi-view pre-training framework to preserve consistency between 2D topological structures and 3D geometric views. (3) Wang et al. (2022a); Fang et al. (2022b) introduce chemical domain knowledge to molecule representations from extra data sources. MolR (Wang et al., 2022a) adopt the composable TransE methods in NLP to preserve the equivalence of molecules with respect to chemical reactions in the embedding space. KCL (Fang et al., 2022b) constructs a chemical element knowledge graph to incorporate prior knowledge into molecular graph semantics. Note that on top of MolR, we incorporate the important factor of chemical reactions—yields—as a measure to adaptively scale the margin of the TrasE loss in the pre-training phase to increase the sample efficiency and obtain more informative molecular representations.

Knowledge Distillation. Knowledge distillation is a generic training paradigm where a student model is trained under the supervision of a teacher model to achieve (implicit) knowledge transfer (Gou et al., 2021). According to Tian et al. (2019), there mainly exist three distillation paradigms:

(1) model compressing (Sanh et al., 2019), (2) cross-modal knowledge transfer (Romero et al., 2014), (3) ensemble distillation (Buciluǎ et al., 2006). We can also classify distillation techniques from the perspective of knowledge categories (Gou et al., 2021): logit-based (Hinton et al., 2015) and feature-based (Tian et al., 2019) distillation method. In this paper, we focus on the cross-modal (from chemical reaction data to molecule data) and feature-based (since it was reported to achieve superior performance than logits-based methods) distillation method. To our best knowledge, this is the first work to consider reaction-to-molecule distillation for molecular property prediction.

3 BACKGROUND

Molecular graphs. Molecules can be naturally represented as graphs by taking atoms as nodes and chemical bonds as edges. Formally, let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{X})$ denote a molecular graph, where $\mathcal{V} = \{v_i\}_{i=1}^{N}$ is a set of N atoms, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of chemical bonds between atoms. $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]^T \in \mathbb{R}^{N \times d}$ represents the atom feature matrix, where d is the feature dimension. Each atom i has the an initial feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$, such as molecule fingerprints.

Chemical reactions. In this paper, a chemical reaction can be described as a pair of molecular graphs with a reaction yield $(\mathcal{G}_R, \mathcal{G}_P, y)$, where $\mathcal{G}_R = \{\mathcal{G}_{r_1}, \mathcal{G}_{r_2}, \cdots\}$ and $\mathcal{G}_P = \{\mathcal{G}_{p_1}, \mathcal{G}_{p_2}, \cdots\}$ are the set of reactants and products respectively, and y is the yield of this reaction. Note that reactants and products of the same reaction always share the same set of nodes because the chemical reaction preserves atom number and types. In organic chemical synthesis processes, the reaction yield is a primary factor, which measures the amount of a specific set of products formed per mole of the reactants consumed (Fogler & Fogler, 1999). Yields are usually expressed as a percentage (ratio of actual yield to theoretical yield) and fall into the range $[0, 1]$. According to Furniss (1989), yields above 70% are relatively good due to side and incomplete reactions that generate other products. In general, the higher yield indicates that the reaction is more efficient and important for organic chemical synthesis in practice.

Uncertainty knowledge graphs (KG). An uncertainty KG consists of a set of weighted triples $\{(l, s_l)\}$ (Chen et al., 2021). For each pair (l, s_l), $l = (h, r, t)$ is a triple representing a relation fact where h is the head entity, t is the tail entity, and r is the associated relation. $s_l \in [0, 1]$ denotes the confidence score for this relation fact to be true. Some examples of weighted triples are as follows (Chen et al., 2019): ((university, *synonym*, institute), 0.86) and ((fork, *atlocation*, kitchen), 0.4). Overall, given the uncertainty KG, we aim to learn an embedding model to encode each entity and relation into a low-dimensional space where confidence scores of relation facts are preserved.

Graph neural networks (GNNs). GNNs have become increasingly popular in various graph mining and molecular modeling tasks (Wu et al., 2020). Typically, the training process of modern GNNs follows the message-passing mechanism (Hamilton, 2020). During each message-passing iteration, a hidden embedding $\boldsymbol{h}_u^{(k)}$ corresponding to each node $u \in \mathcal{V}$ can be expressed as follows:

$$\begin{aligned} \boldsymbol{h}_u^{(k)} &= \text{UPDATE}^{(k)}\left(\boldsymbol{h}_u^{(k-1)}, \text{AGGREGATE}^{(k)}\left(\left\{\boldsymbol{h}_v^{(k-1)}, v \in \mathcal{N}(u)\right\}\right)\right) \\ &= \text{UPDATE}^{(k)}\left(\boldsymbol{h}_u^{(k-1)}, \boldsymbol{m}_{\mathcal{N}(u)}^{(k)}\right), \end{aligned} \tag{1}$$

where UPDATE and AGGREGATE are arbitrary differentiable functions (*i.e.*, neural networks), and $\boldsymbol{m}_{\mathcal{N}(u)}$ denotes the "message" that is aggregated from u's neighborhood $\mathcal{N}(u)$. The initial embedding at $k = 0$ is set to the input features, *i.e.*, $\boldsymbol{h}_u^{(0)} = x_u, \forall u \in \mathcal{V}$. After running k iterations of the GNN message-passing, we can obtain information from k-hops neighborhood nodes. Different GNNs can be obtained by choosing different UPDATE and AGGREGATE functions. For graph classification tasks, a graph-level representation $\boldsymbol{h}_{\mathcal{G}}$ is obtained by integrating all the node embeddings $\boldsymbol{h}_u^{(K)}$ among the graph $\mathcal{G}$ after K iterations:

$$\boldsymbol{h}_{\mathcal{G}} = \text{READOUT}\left(\left\{\boldsymbol{h}_u^{(K)}, u \in \mathcal{V}\right\}\right), \tag{2}$$

where the READOUT$(\cdot)$ is a permutation invariant function such as summation and maximization operators or more complex pooling methods (Ying et al., 2018; Gao & Ji, 2019).

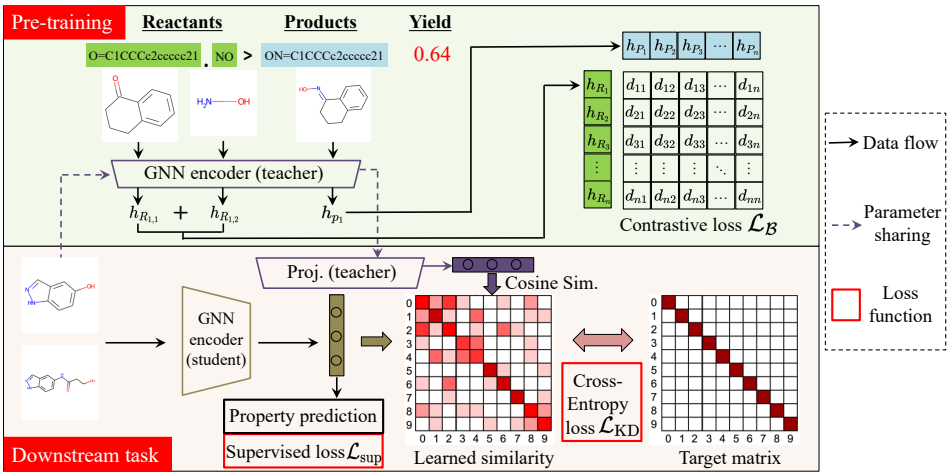

Figure 2: An overview of our MolKD framework. MolKD consists of two-stage training phases: the yield-guided chemical reaction pre-training and the reaction-to-molecule distillation.

Problem setup. For molecular property prediction, given a training set of molecular graphs $\{\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_{N_l}\} \subseteq \mathbb{G}$ and their labels $\{l_1, l_2, \cdots, l_{N_l}\} \subseteq \mathbb{L}$, *e.g.*, quantum mechanics property, our goal is to learn a function $f : \mathbb{G} \to \mathbb{L}$ and predict the labels of other molecular graphs in the testing set. In this paper, the student encoder f^S of the learned function f is supervised by our proposed pre-trained teacher model f^T, which transfers knowledge from chemical reaction data.

4 MolKD: The Proposed Method

In this section, we present our MolKD method according to its two-stage training phases. We first introduce how to preserve equivalence of chemical reactions and incorporate information of yields in the pre-training phase (Sec. 4.1). After that, we propose the reaction-to-molecule distillation to transfer cross-modal knowledge between pre-training reaction data and the downstream molecule data (Sec. 4.2). The schematic illustration of our proposed method is shown in Fig. 2.

4.1 Yield-guided Chemical Reaction Pre-training

Given the chemical reaction dataset, we design the self-supervised learning task for pre-training to obtain informative molecular representations. For the sake of intuitive representation, we use the condensation reaction of methyl acetate $(\mathrm{CH_3COOH} + \mathrm{C_2H_5OH} \longrightarrow \mathrm{CH_3COOCH_3} + \mathrm{H_2O}, 0.9)$ [2] as the representative chemical reaction for illustration in what follows.

Encoding molecular embeddings. In order to encode rich structure information of the given molecule, we first convert molecular SMILES (Weininger, 1988) strings to molecular graphs. We utilize GNNs to encode molecular graphs as numeric vectors. Zhang et al. (2021) argued that hydrogen atoms can help determine the number of the chemical bonds for atoms. Therefore we regard hydrogen atoms as independent nodes in molecular graphs. We then utilize PySmiles (Landrum, 2013) library to produce six types of atom properties: *element type*, *(anti)-aromaticness*, *mass*, *the number of implicit hydrogens*, *charge*, and *class*. Each type of atomic property is represented as a one-hot vector. The initial node feature of each atom in the molecular graph is the concatenation of six one-hot vectors. Following Wang et al. (2022a), we do not explicitly encode edge attributes because they can be implicitly inferred by the node features of their own corresponding atoms. Intuitively, we take ethanol $\mathrm{C_2H_5OH}$ as input, we can get its representation vector $h_{\mathrm{C_2H_5OH}}$ after this step.

[2] Chemical reactions usually occurs under some specific conditions, such as catalysts. The complete equation of this chemical reaction is $\left(\mathrm{CH_3COOH} + \mathrm{C_2H_5OH} \xrightarrow[\Delta]{\mathrm{H_2SO_4}} \mathrm{CH_3COOCH_3} + \mathrm{H_2O}, 0.9\right)$. Following previous work Wang et al. (2022a), we omit the chemical conditions for clarity because they do not affect the conservation law (*e.g.*, atom number and types) between reactants and products.

Preserving chemical reaction equivalence. A chemical reaction defines a transformation relation "$\rightarrow$" between the reactant set $R = \{r_1, r_2, \cdots, r_n\}$ and the product set $P = \{p_1, p_2, \cdots, p_m\}$:

$$r_1 + r_2 + \cdots + r_n \longrightarrow p_1 + p_2 + \cdots + p_m. \tag{3}$$

A chemical reaction represents a particular relation of *equivalence* between its reactants and products in terms of the conservation of mass and charge. We adopt the idea of preserving the equivalence of chemical reactions in molecular embedding space (Wang et al., 2022a). This *equivalence* property is reminiscent of the translation-based model in KG that learns embeddings by narrowing the distance between the head and tail entity transformed by the relation to preserve its composable property. If we take reactants of a reaction as the head entity and products as the tail entity, we can impose constraints on molecular embeddings to preserve the *equivalence* property [3]:

$$\sum_{r \in R} h_r = \sum_{p \in P} h_p. \tag{4}$$

For a minibatch data $\mathcal{B}_1 = \{(R_1, P_1, y_1), (R_2, P_2, y_2), \cdots\}$, we compute a score function to measure the fitness of the reactant-product pair in a reaction. We adopt the L-2 norm introduced in TransE (Bordes et al., 2013) to compute the score function with $f(R, P) = \left\| \sum_{r \in R} h_r - \sum_{p \in P} h_p \right\|_2$. To prevent the TransE model from learning a trivial solution where all molecular representations are equal to zero, we use the contrastive learning strategy (Jaiswal et al., 2020) by drawing unpaired reactants and products as negative samples. The margin-loss function to be minimized is denoted as

$$\mathcal{L}_{\mathcal{B}} = \frac{1}{|\mathcal{B}_1|(|\mathcal{B}_1| - 1)} \sum_i \sum_{i \neq j} \max \left(\left\| \sum_{r \in R_i} h_r - \sum_{p \in P_i} h_p \right\|_2 - \left\| \sum_{r \in R_i} h_r - \sum_{p \in P_j} h_p \right\|_2 + \gamma, 0 \right), \tag{5}$$

where $\gamma > 0$ is a margin hyper-parameter. See Fig. 2 for an illustration. Intuitively, in the molecular embedding space, we obtain $h_{\mathrm{CH_3COOH}} + h_{\mathrm{C_2H_5OH}} \approx h_{\mathrm{CH_3COOCH_3}} + h_{\mathrm{H_2O}}$ for the reaction $\mathrm{CH_3COOH} + \mathrm{C_2H_5OH} \longrightarrow \mathrm{CH_3COOCH_3} + \mathrm{H_2O}$ after this step.

Modeling transformation efficiency with reaction yields. The yield is a crucial measure of transformation efficiency of the reactant-product pair in a reaction. Intuitively, the higher yield of a reaction suggests that given its reactants, it's more likely to generate the corresponding products than others, *i.e.*, there is higher probability that the relation between the reactants and products is true. This draws a conceptual analogy to the confidence score on an uncertainty KG. Accordingly, we adopt the idea of GTransE (Kertkeidkachorn et al., 2019) to capture the confidence of reaction triples in an uncertainty KG, by adaptively modifying the margin with the reaction yield in pre-training. Intuitively, the higher the yield, the larger margin should be set to pay more attention to this reaction. Thus, the margin γ should be increased when the reaction yield y gets higher. In Fig. 3, we maximize the margin between the positive sample $f^+(R_i, P_i) = \left\| \sum_{r \in R_i} h_r - \sum_{p \in P_i} h_p \right\|_2$ and the negative sample $f^-(R_i, P_j) = \left\| \sum_{r \in R_i} h_r - \sum_{p \in P_j} h_p \right\|_2$ of the higher yield (the red rectangle) rather than the lower yield one (the green rectangle). Formally, if we have (R_1, P_1, y_1), (R_2, P_2, y_2), and $y_1 > y_2$, then $f^+(R_1, P_1) - f^-(R_1, P_2) > f^+(R_2, P_2) - f^-(R_2, P_1)$. Therefore, the margin-loss function of yield-guided chemical reaction pre-training can be further derived as [4]

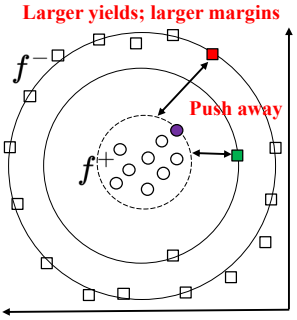

Figure 3: The illustration of the embedding space with different margins according to reaction yields. Circles and rectangles denote positive and negative samples.

$$\mathcal{L}_{\mathcal{B}} = \frac{1}{|\mathcal{B}_1|(|\mathcal{B}_1| - 1)} \sum_i \sum_{i \neq j} \max \left(\left\| \sum_{r \in R_i} h_r - \sum_{p \in P_i} h_p \right\|_2 - \left\| \sum_{r \in R_i} h_r - \sum_{p \in P_j} h_p \right\|_2 + y_i^{\alpha} \gamma, 0 \right), \tag{6}$$

[3] We omit the modeling of relations because there exists only one transformation relation "$\rightarrow$" in a chemical reaction. We leave the modeling of different chemical reaction types and conditions as future work.

[4] Here we only replace the corrupted products P in negative samples because yields reflect the transformation efficiency of the main product in a reaction.

where $\alpha \geq 0$ is a hyper-parameter to control the influence of reaction yields. As α becomes larger, the effect of yields on model transformation efficiency is amplified. When α equals zero, the influence of uncertainty is eliminated, and Eq. 6 becomes Eq. 5. Intuitively, in the molecule embedding space, we obtain $h_{\mathrm{CH_3COOH}} + h_{\mathrm{C_2H_5OH}} \approx h_{\mathrm{CH_3COOCH_3}} + h_{\mathrm{H_2O}}$ for $(\mathrm{CH_3COOH} + \mathrm{C_2H_5OH} \xrightarrow[\Delta]{\mathrm{H_2SO_4}} \mathrm{CH_3COOCH_3} + \mathrm{H_2O}, 0.9)$ considering its reaction yield 0.9 after this step.

4.2 Reaction-to-molecule Distillation

Given molecule data with property labels, we train a predictor supervised by the pre-trained model. Here we introduce the feature-based knowledge distillation method to transfer knowledge in chemical reaction data for molecular property prediction tasks. The workflow of MolKD is illustrated in Fig. 2. Again, we use the toxicity of ethanol $\mathrm{C_2H_5OH}$ as the representative molecular property prediction task for illustration in what follows.

We adopt the feature-based distillation method, *i.e.*, contrastive representation distillation (Tian et al., 2019) and formulate representation distillation as a contrastive learning task on pairwise relationships between the teacher and student representations. Intuitively, we want to maximize the consistency of representations for the teacher and student model. Given data of molecular property prediction $\mathcal{B}_2 = \{(\mathcal{G}_1, l_1), (\mathcal{G}_2, l_2), \cdots\}$, we maximize the similarity among pairs of student and teacher representations corresponding to the same molecular graph, *i.e.*, $f^S(\mathcal{G}_i), f^T(\mathcal{G}_i)$ (positive samples), while pushing away the representations of pairs of unmatched molecules, *i.e.*, $f^S(\mathcal{G}_i), f^T(\mathcal{G}_j)$ (negative samples). We utilize the InfoNCE loss (Oord et al., 2018) in reaction-to-molecule distillation. Given two molecular encoders f^S and f^T with the same output feature dimension, we denote representation distillation as the task of classifying positive pairs among a set of negative pairs as

$$\mathcal{L}_{\mathrm{KD}}(f^S, f^T) = -\frac{1}{|\mathcal{B}_2|} \sum_i \log \frac{\exp\left(s\left(f^S(\mathcal{G}_i), f^T(\mathcal{G}_i)\right)/\tau\right)}{\exp\left(s\left(f^S(\mathcal{G}_i), f^T(\mathcal{G}_i)\right)/\tau\right) + \sum_{j \neq i} \exp\left(s\left(f^S(\mathcal{G}_i), f^T(\mathcal{G}_j)\right)/\tau\right)}, \tag{7}$$

where τ represents the temperature hyper-parameter and $s(\cdot)$ indicates the similarity between molecular representations. Here we use the cosine similarity function. Intuitively, we fine-tune the student model f^S to predict the toxicity of $\mathrm{C_2H_5OH}$ supervised by the teacher model f^T pre-trained on reaction $(h_{\mathrm{CH_3COOH}} + h_{\mathrm{C_2H_5OH}} \approx h_{\mathrm{CH_3COOCH_3}} + h_{\mathrm{H_2O}}, 0.9)$.

Training objective. In MolKD, we use TAG (Du et al., 2017) as our backbone GNN model. As shown in Fig. 2, we need to optimize two components of the loss function. First, we have an auxiliary reaction-to-molecule distillation loss to obtain a powerful GNN encoder for effective molecular representations (Eq. 6). Second, given the training set of molecular data $(\mathcal{G}, l) \in \mathbb{G} \times \mathbb{L}$, the GNN model f parameterized by $\boldsymbol{\theta}$ is further optimized with a supervised loss $\mathcal{L}_{\text{sup}}$ as follows:

$$\mathcal{L}_{\text{sup}} = \mathbb{E}_{(\mathcal{G},l)\in\mathbb{G}\times\mathbb{L}}[\mathcal{L}\left(f_{\boldsymbol{\theta}}\left(\mathcal{G}\right), l\right)]\ . \tag{8}$$

Overall, the final training objective of MolKD is

$$\mathcal{L} = \beta \cdot \mathcal{L}_{\text{sup}} + (1-\beta) \cdot \mathcal{L}_{\mathrm{KD}}\left(f^S, \Phi(f^T)\right), \tag{9}$$

where β is a scaling factor to balance the supervised loss and the representation distillation loss, and Φ is an MLP layer to transform the feature maps of the teacher and student models in the same shape.

5 Experiments

In this section, we conduct extensive experiments to answer the following questions: (1) Can our proposed yield-guided chemical reaction pre-training method MolKD provide effective molecular representations? (Sec. 5.1) (2) How does MolKD compare to state-of-the-art methods on molecular property prediction tasks? (Sec. 5.2) (3) Can MolKD serve as a reliable method with high robustness and interpretability? (Sec. 5.3) Each result in this section is obtained by running over 10 runs, and we show the results of mean and standard deviation on the testing set. Due to space limitations, the descriptions and statistics of the datasets we use are provided in Appendix A. Additional empirical studies such as ablations of the backbone GNN models and knowledge distillation methods, as well as the results of regression tasks on molecular property prediction tasks, can be found in Appendix B.

Table 1: Results of chemical reaction prediction on USPTO-500-MT. We use the pre-trained models of Mol2Vec and MolBERT from their respective papers. MolR and MolKD variants are pre-trained on the training set of USPTO-500-MT. Arrows indicate the direction of better performance. **Bold** / underline denote the best / second-best result for each column.

	MRR ↑	MR ↓	Hit@1 ↑	Hit@3 ↑	Hit@5 ↑	Hit@10 ↑
Mol2Vec	0.758	93.748	0.696	0.799	0.830	0.864
MolBERT	0.796	18.781	0.720	0.853	0.888	0.923
MolR	$0.853_{\pm 0.005}$	$2.351_{\pm 0.108}$	$0.783_{\pm 0.006}$	$0.911_{\pm 0.005}$	$0.941_{\pm 0.004}$	$0.968_{\pm 0.002}$
MolKD-random	$0.725_{\pm 0.017}$	$174.740_{\pm 49.859}$	$0.669_{\pm 0.021}$	$0.762_{\pm 0.015}$	$0.790_{\pm 0.013}$	$0.822_{\pm 0.011}$
MolKD-confidence	$\underline{0.858_{\pm 0.005}}$	$2.271_{\pm 0.106}$	$\underline{0.789_{\pm 0.007}}$	$\underline{0.915_{\pm 0.005}}$	$\underline{0.945_{\pm 0.004}}$	$\underline{0.970_{\pm 0.002}}$
MolKD-CKRL	$0.857_{\pm 0.005}$	$\underline{2.260_{\pm 0.099}}$	$0.787_{\pm 0.007}$	$0.913_{\pm 0.005}$	$0.944_{\pm 0.003}$	$\underline{0.970_{\pm 0.002}}$
MolKD	$\mathbf{0.875_{\pm 0.003}}$	$\mathbf{2.058_{\pm 0.054}}$	$\mathbf{0.811_{\pm 0.003}}$	$\mathbf{0.929_{\pm 0.002}}$	$\mathbf{0.955_{\pm 0.002}}$	$\mathbf{0.977_{\pm 0.002}}$

5.1 Pre-training on Chemical Reactions

Datasets. We use USPTO-500-MT collected by Lu & Zhang (2022), a public large-scale reaction dataset including reaction SMILES and yields, to validate the effectiveness of the yield-guided chemical reaction pre-training method. This dataset consists of 143,535 reactions: 116,360/12,937/14,238 for training/validation/testing. Each reaction has at most 5 reactants and only 1 main product.

Baselines. We compare MolKD with various pre-training methods for molecular representations: Mol2vec (Jaeger et al., 2018), MolBERT (Fabian et al., 2020), and MolR (Wang et al., 2022a). Moreover, we carry out the following ablations: (1) we replace reaction yields with random numbers falling into $[0, 1]$ (MolKD-random), and confidence scores (Ghiandoni et al., 2019) that measure the quality of reaction classes (MolKD-confidence). These two ablations are designed to highlight the importance of yields when pre-training on reactions. (2) we adopt another uncertainty KG method, *i.e.*, CKRL (Xie et al., 2018)) to validate the versatility of our model (MolKD-CKRL).

Setups. Inspired by the equivalence property between reactants and products that $h_R \approx h_P$, we treat the chemical reaction prediction as a ranking problem following Wang et al. (2022a). In the testing phase, given the reactants of a reaction, we predict its main product and rank all candidate products (there are 14,123 in total) according to l-2 distances in the representation space $||h_R - h_p||_2$. We conduct three measures as our evaluation metrics. (1) MRR: mean reciprocal rank (2) MR: mean rank of correct entities (3) Hit@K: the proportion of correct answers ranked in top K.

Table 1 shows the results of chemical reaction predictions, from which we can observe that: (1) Our MolKD model achieves the best performance compared with other baselines and ablations over 6 evaluation metrics. For example, MolKD achieves 2.2% absolute MRR gain and 2.8% absolute Hit@1 gain compared with the best-performed MolR model, which confirms the capability of our yield-guided pre-training method. (2) Compared with MolR (without considering yields), MolKD-random (replacing yields with meaningless numbers), and MolKD-confidence (replacing yields with reaction confidence scores (Ghiandoni et al., 2019)), MolKD achieves $2.2\%/12.0\%/1.7\%$ absolute MRR improvement on MRR, highlighting the effectiveness of introducing the important chemical reaction factor–yields when pre-training molecular representations. (3) MolKD achieves slightly better performance compared with MolKD-CKRL, which suggests that MolKD does not rely on the specific uncertainty KG method to achieve strong performance.

Investigation of pre-training representations. To examine the effectiveness of the representations learned by our pre-trained MolKD, we visualize a representative molecule with its eight closest molecules in the representation domain. Specifically, given the query molecule (PubChem ID 17842486), we obtain its representations via the yield-guided pre-training method in MolKD and calculate cosine distances with all reference molecules in our pre-training dataset. The cosine distance between two molecular representations $(\boldsymbol{u}, \boldsymbol{v})$ is defined as $1 - \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{||\boldsymbol{u}||||\boldsymbol{v}||}$. Then, all reference molecules are ranked by cosine distances. In Fig. 4, we show the eight closest molecules compared to the query molecule. We also calculate molecular similarities through Tanimoto coefficient (Bajusz et al., 2015) between the query and the selected molecule, and SA scores (Ertl & Schuffenhauer, 2009) to assess the associated molecule's synthetic complexity. We find that these molecules are structurally similar to the query molecule with the same functional groups (*e.g.*, aromatics). They have high Tanimoto similarities larger than 0.5. The learned representations via MolKD are in line with our expectation that molecules with similar structures tend to be close in the representation domain.

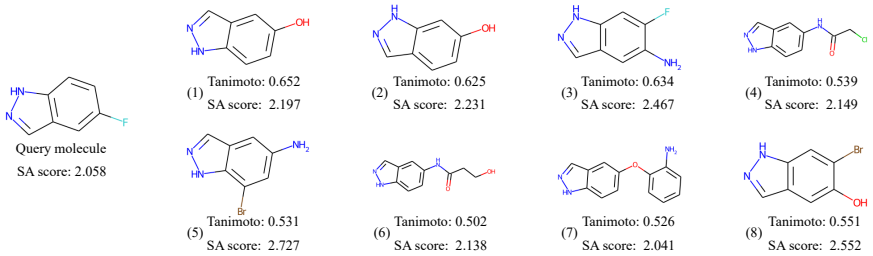

Figure 4: Comparisons of the query molecule (PubChem ID 17842486) and eight closest molecules in MolKD representation domain with Tanimoto similarities and SA score labeled.

Moreover, these molecules have similar SA scores, which further suggests that selected molecules share similar synthetic complexities in absence of the activity cliff issue (Hu & Bajorath, 2012).

5.2 Molecular Property Prediction

Datasets. We benchmark the performance of MolKD on multiple challenging classification and regression tasks from MoleculeNet (Wu et al., 2018). Following Wang et al. (2022a), all datasets are randomly split into training, validation, and testing by 8:1:1. We use BACE, BBBP, ClinTox, HIV, SIDER, and Tox21 for classification tasks and adopt AUC-ROC as the evaluation metric for these binary prediction tasks, for which higher is better. We use ESOL, FreeSolv, and QM8 for regression tasks and adopt the root mean square error (RMSE) for ESOL and FreeSolv, whereas we use mean average error (MAE) for QM8, for which lower is better.

Baselines. We compare MolKD with multiple competitive baselines. AttentiveFP (Xiong et al., 2019), GCN (Kipf & Welling, 2016), D-MPNN (Yang et al., 2019), SchNet (Schütt et al., 2018) are the GNN-based methods without pre-training. As introduced in Sec. 2, the following pre-training methods are divided into three categories. (1) Mol2vec (Jaeger et al., 2018), ChemBERTa (Chithrananda et al., 2020), MolBert (Fabian et al., 2020), PretrainGNN (Hu et al., 2020), GROVER-base (Rong et al., 2020), and MolCLR (Wang et al., 2022b) with carefully designed self-supervised learning tasks. (2) GEM (Fang et al., 2022a) and GraphMVP (Liu et al., 2022a) with 2D-3D geometric alignment. (3) KCL (Fang et al., 2022b) and MolR (Wang et al., 2022b) with auxiliary chemical data sources. MolKD$^-$ means we directly use the pre-trained GNN model followed by a logistic regression layer to make predictions, which does not use the reaction-to-molecule distillation method.

The overall performance of MolKD along with other baselines on classification tasks are summarized in Table 2. We have the following observations: (1) MolKD achieves the state-of-the-art performance over 5 out of 6 classification tasks, which demonstrates the effectiveness of two main components in MolKD: the yield-guided pre-training and the reaction-to-molecule distillation. For example, MolKD achieves [8.0%, 1.6%] absolute AUC-ROC gain on Tox21 compared with the best-performed method [AttentiveFP, KCL] ([w, w/o pre-training]). (2) Methods with pre-training on large-scale unlabeled molecules consistently outperform methods without pre-training. Notably, MolKD$^-$ achieves the on par performance compared with MolR over all datasets. However, the size of pre-training data in MolKD is almost 1/3 of that in MolR. These results highlight the importance of introducing yields to increase data efficiency when pre-training on reactions. (3) MolKD achieves the best performance with a minimal number of model parameters, which further corroborates its effectiveness.

5.3 Experimental Analysis

Robustness against molecular structure perturbation. To evaluate the robustness of our proposed method, we follow the principle of property-slightly-affected structure perturbation (PASP) introduced by Li et al. (2022b) and utilize the PhysProp dataset to perform a robustness experiment. The aim is to investigate that the model does not significantly affect the predictions (molecular properties) when the input perturbed molecule set suffers small perturbations. More details of the evaluation metric (*i.e.*, effect

Table 3: Effect score of molecular structure perturbation test.

Method	Effect score$_{std}$ [lower is better]		
	Level1	Level2	Level3
GCN	$0.385_{\pm 0.161}$	$0.712_{\pm 0.169}$	$0.997_{\pm 0.183}$
GAT	$0.388_{\pm 0.055}$	$0.615_{\pm 0.087}$	$0.943_{\pm 0.145}$
GIN	$0.312_{\pm 0.017}$	$0.526_{\pm 0.039}$	$0.764_{\pm 0.015}$
MPNN	$0.315_{\pm 0.014}$	$0.518_{\pm 0.054}$	$0.750_{\pm 0.048}$
GLAM	$\underline{0.290_{\pm 0.010}}$	$\underline{0.493_{\pm 0.074}}$	$\mathbf{0.656_{\pm 0.118}}$
MolKD	$\mathbf{0.251_{\pm 0.021}}$	$\mathbf{0.491_{\pm 0.036}}$	$\underline{0.710_{\pm 0.035}}$

Table 2: Results of molecular property prediction on classification tasks (metric: AUC-ROC). The first block is the conventional methods without pre-training thus we omit their *#Pre-training data*. We split the pre-training methods in the second to the fourth block according to three method categories introduced in Sec. 2. The results in the first four blocks are taken from their original papers and "-" means they do not report the corresponding results. "Z", "C", "P", "Gu", "Ge", 'U-4", and "U-1" in the *#Pre-training data* column denote the used pre-training data, which are the abbreviation of ZINC15 (Sterling & Irwin, 2015), ChEMBL (Gaulton et al., 2012), PubChem (Kim et al., 2019), GuacaMol (Brown et al., 2019), GEOM (Axelrod & Gomez-Bombarelli, 2022), USPTO-479k (Wang et al., 2022a), and USPTO-500-MT (140k) (Lu & Zhang, 2022) respectively.

	BACE	BBBP	ClinTox	HIV	SIDER	Tox21	#Params (M)	#Pre-training data (M)
AttentiveFP	0.784	0.643	0.847	0.771	0.606	0.761	-	
GCN	0.716	0.718	0.625	0.740	0.536	0.709	1.04	*
D-MPNN	-	0.708	0.906	0.752	0.632	0.688	1.41	
SchNet	$0.766_{\pm 0.011}$	$0.848_{\pm 0.022}$	$0.715_{\pm 0.037}$	$0.702_{\pm 0.034}$	$0.539_{\pm 0.037}$	$0.727_{\pm 0.023}$	2.20	
Mol2vec	$0.862_{\pm 0.027}$	$0.872_{\pm 0.021}$	$0.841_{\pm 0.062}$	$0.769_{\pm 0.021}$	-	$0.803_{\pm 0.041}$	4.87	19.90 (Z+C)
ChemBERTa	-	0.643	0.733	0.622	-	0.728	56.44	77.00 (P)
MolBert	0.866	0.762	-	0.783	-	0.806	64.42	1.60 (Gu)
PretrainGNN	$0.845_{\pm 0.007}$	$0.687_{\pm 0.013}$	$0.726_{\pm 0.015}$	$0.799_{\pm 0.007}$	$0.627_{\pm 0.008}$	$0.781_{\pm 0.006}$	4.36	2.00 (Z)
GROVER-base	$0.878_{\pm 0.016}$	$0.936_{\pm 0.008}$	$0.925_{\pm 0.013}$	-	$0.656_{\pm 0.006}$	$0.819_{\pm 0.020}$	48.39	11.00 (Z+C)
MolCLR-GCN	$0.788_{\pm 0.005}$	$0.738_{\pm 0.002}$	$0.867_{\pm 0.010}$	$0.778_{\pm 0.005}$	$0.669_{\pm 0.012}$	$0.747_{\pm 0.008}$	1.04	10.00 (P)
MolCLR-GIN	$\mathbf{0.890_{\pm 0.003}}$	$0.736_{\pm 0.005}$	$0.932_{\pm 0.017}$	$0.806_{\pm 0.011}$	$0.680_{\pm 0.011}$	$0.798_{\pm 0.007}$	2.40	10.00 (P)
GEM	$0.856_{\pm 0.011}$	$0.724_{\pm 0.004}$	$0.901_{\pm 0.013}$	$0.806_{\pm 0.009}$	$0.672_{\pm 0.004}$	$0.781_{\pm 0.001}$	1.66	20.00 (Z)
GraphMVP	$0.768_{\pm 0.011}$	$0.685_{\pm 0.002}$	$0.790_{\pm 0.025}$	$0.748_{\pm 0.014}$	$0.623_{\pm 0.016}$	$0.745_{\pm 0.004}$	3.22	0.05 (GE)
KCL	0.860	0.927	0.898	-	0.659	0.825	1.16	0.25 (Z)
MolR-GCN	$0.882_{\pm 0.019}$	$0.890_{\pm 0.032}$	$0.916_{\pm 0.0394}$	$0.802_{\pm 0.024}$	-	$0.818_{\pm 0.023}$	1.98	0.48 (U-4)
MolR-TAG	$0.875_{\pm 0.023}$	$0.895_{\pm 0.031}$	$0.913_{\pm 0.043}$	$0.801_{\pm 0.023}$	-	$0.820_{\pm 0.028}$	3.26	0.48 (U-4)
MolKD^{-}	$0.876_{\pm 0.031}$	$0.902_{\pm 0.028}$	$0.917_{\pm 0.039}$	$0.811_{\pm 0.021}$	$0.660_{\pm 0.048}$	$0.818_{\pm 0.059}$	3.26	0.14 (U-1)
MolKD (ours)	$0.872_{\pm 0.022}$	$\mathbf{0.942_{\pm 0.021}}$	$\mathbf{0.933_{\pm 0.041}}$	$\mathbf{0.816_{\pm 0.011}}$	$\mathbf{0.706_{\pm 0.020}}$	$\mathbf{0.841_{\pm 0.032}}$	0.84	0.14 (U-1)

score) are provided in Appendix A. Table 3 shows that MolKD achieves the best performance with high robustness on the level 1&2 perturbations and is less affected by molecular structure perturbations. We postulate the reason is that pre-training on reactions could incorporate comprehensive chemical domain knowledge to increase the robustness of the predictor.

Interpretability. To better understand the predictors generated by MolKD, we investigated its decision-making process and interpreted its learned knowledge on PhysProp (Li et al., 2022b). As shown in Fig. 5, we visualize some case studies of the solubility prediction and explain the model from the hidden states of the last layer by averaging scaling. In general, hydroxyl and amino groups are considered to be more hydrophilic, and alkyl and halogen groups are considered to be more lipophilic. In Fig. 5, the atoms in the hydrophilic group (*e.g.*, -OH, and -COOH) tend to be bluer and their weights are closer to 1 in our visualization. Meanwhile, the atoms in the lipophilic group (*e.g.*, benzene ring) tend to be redder and their weights are closer to -1. These results are in line with the chemical intuition, which suggests that MolKD captures interpretable knowledge and patterns in chemistry. More case studies of interpretability can be found in Appendix B.

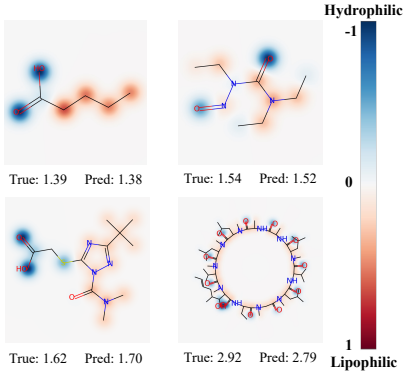

Figure 5: Case studies of atom-level interpretation with true and predicted solubility labeled.

6 Conclusion

In this work, we study how to distill knowledge in chemical reactions to assist molecular property prediction. We propose a novel method named MolKD to obtain effective molecular representations, which consists of two main components: the yield-guided chemical reaction pre-training method and the reaction-to-molecule distillation. Through extensive experiments on chemical reaction prediction and molecular property prediction, we show that MolKD achieves significantly superior performance with high robustness and interpretability. We hope our work could stimulate more ideas to squeeze the potential of chemical domain knowledge for molecular property prediction.

Reproducibility statement. The source code along with a README file with instructions on how to run these experiments is attached in the supplementary material. In addition, more discussions about dataset descriptions and statistics, hyper-parameters for our proposed model, and experimental settings are detailed in Appendix A.

Ethics statement. In this work, our studies are not related to human subjects, practices to data set releases, potentially harmful insights, discrimination/bias/fairness concerns, and also do not have legal compliance or research integrity issues. All datasets we used are curated from public data sources, and are released under a license that allowed for public access. The datasets are all anonymized. We develop a novel and practical method to incorporate chemical domain knowledge for popular molecular property prediction problems. Therefore, we do not foresee any particular concerns related to its ethical aspects or future societal consequences. However, advanced computational chemistry tools may pose the risk of misuse, *e.g.*, for the development of chemical weapons. Our studies, to the best of our knowledge, do not promote misuse any more than computational chemistry research.

REFERENCES

Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. *Nature Machine Intelligence*, 3(12):1023–1032, 2021.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):1–14, 2022.

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

Hangrui Bi, Hengyi Wang, Chence Shi, Connor Coley, Jian Tang, and Hongyu Guo. Non-autoregressive electron redistribution modeling for reaction prediction. In *International Conference on Machine Learning*, pp. 904–913. PMLR, 2021.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3): 1096–1108, 2019.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Xuelu Chen, Muhao Chen, Weijia Shi, Yizhou Sun, and Carlo Zaniolo. Embedding uncertain knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3363–3370, 2019.

Xuelu Chen, Michael Boratko, Muhao Chen, Shib Sankar Dasgupta, Xiang Lorraine Li, and Andrew McCallum. Probabilistic box embeddings for uncertain knowledge graph reasoning. *arXiv preprint arXiv:2104.04597*, 2021.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Jian Du, Shanghang Zhang, Guanhang Wu, José MF Moura, and Soummya Kar. Topology adaptive graph convolutional networks. *arXiv preprint arXiv:1710.10370*, 2017.

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

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.

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

Yin Fang, Qiang Zhang, Haihong Yang, Xiang Zhuang, Shumin Deng, Wen Zhang, Ming Qin, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Molecular contrastive learning with chemical element knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3968–3976, 2022b.

H Scott Fogler and Scott H Fogler. *Elements of chemical reaction engineering*. Pearson Educacion, 1999.

Brian S Furniss. *Vogel's textbook of practical organic chemistry*. Pearson Education India, 1989.

Hongyang Gao and Shuiwang Ji. Graph u-nets. In *international conference on machine learning*, pp. 2083–2092. PMLR, 2019.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Gian Marco Ghiandoni, Michael J Bodkin, Beining Chen, Dimitar Hristozov, James EA Wallace, James Webster, and Valerie J Gillet. Development and application of a data-driven reaction classification model: comparison of an electronic lab notebook and medicinal chemistry literature. *Journal of chemical information and modeling*, 59(10):4167–4187, 2019.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

William L Hamilton. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning*, 14(3):1–159, 2020.

Zhongkai Hao, Chengqiang Lu, Zhenya Huang, Hao Wang, Zheyuan Hu, Qi Liu, Enhong Chen, and Cheekong Lee. Asgn: An active semi-supervised graph neural network for molecular property prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 731–752, 2020.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1921–1930, 2019.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Ye Hu and Jürgen Bajorath. Extending the activity cliff concept: structural categorization of activity cliffs and systematic identification of different types of cliffs in the chembl database. *Journal of chemical information and modeling*, 52(7):1806–1811, 2012.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

Sabrina Jaeger, Simone Fulle, and Samo Turk. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58(1):27–35, 2018.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Natthawut Kertkeidkachorn, Xin Liu, and Ryutaro Ichise. Gtranse: generalizing translation-based model on uncertain knowledge graph embedding. In *Annual Conference of the Japanese Society for Artificial Intelligence*, pp. 170–178. Springer, 2019.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

Pengyong Li, Jun Wang, Yixuan Qiao, Hao Chen, Yihuan Yu, Xiaojun Yao, Peng Gao, Guotong Xie, and Sen Song. Learn molecular representations from large-scale unlabeled molecules for drug discovery. *arXiv preprint arXiv:2012.11175*, 2020.

Shuangli Li, Jingbo Zhou, Tong Xu, Dejing Dou, and Hui Xiong. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4541–4549, 2022a.

Yuquan Li, Chang-Yu Hsieh, Ruiqiang Lu, Xiaoqing Gong, Xiaorui Wang, Shuo Liu, Yanan Tian, Dejun Jiang, Jiaxian Yan, Qifeng Bai, et al. An adaptive graph learning method for automated molecular interactions and properties predictions. In *Nat Mach Intell (2022).*, 2022b.

Shengchao Liu, Meng Qu, Zuobai Zhang, Huiyu Cai, and Jian Tang. Structured multi-task learning for molecular property prediction. In *International Conference on Artificial Intelligence and Statistics*, pp. 8906–8920. PMLR, 2022a.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations (ICLR)*, 2022b.

Jieyu Lu and Yingkai Zhang. Unified deep learning model for multitask reaction predictions with explanation. *Journal of Chemical Information and Modeling*, 62(6):1376–1387, 2022.

Youzhi Luo and Shuiwang Ji. An autoregressive flow model for 3d molecular geometry generation from scratch. In *International Conference on Learning Representations*, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

Wan Xiang Shen, Xian Zeng, Feng Zhu, Chu Qin, Ying Tan, Yu Yang Jiang, Yu Zong Chen, et al. Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nature Machine Intelligence*, 3(4):334–343, 2021.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, pp. 20479–20502. PMLR, 2022.

Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.

Igor V Tetko, Vsevolod Yu Tanchuk, and Alessandro EP Villa. Prediction of n-octanol/water partition coefficients from physprop database using artificial neural networks and e-state indices. *Journal of chemical information and computer sciences*, 41(5):1407–1421, 2001.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin D Burke. Chemical-reaction-aware molecule representation learning. In *International Conference on Learning Representations (ICLR)*, 2022a.

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

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

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

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 2020.

Ruobing Xie, Zhiyuan Liu, Fen Lin, and Leyu Lin. Does william shakespeare really write hamlet? knowledge representation learning with confidence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

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

Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *Advances in neural information processing systems*, pp. 4805–4815, 2018.

Xiao-Chen Zhang, Cheng-Kun Wu, Zhi-Jiang Yang, Zhen-Xing Wu, Jia-Cai Yi, Chang-Yu Hsieh, Ting-Jun Hou, and Dong-Sheng Cao. Mg-bert: leveraging unsupervised atomic representation learning for molecular property prediction. *Briefings in bioinformatics*, 22(6):bbab152, 2021.

A EXPERIMENTAL DETAILS

In this section, we first provide the detailed descriptions and statistics of all datasets used in Sec. 5. We then list hyper-parameters of our proposed MolKD method followed by the experimental settings.

A.1 DATASET DESCRIPTIONS AND STATISTICS

To benchmark the performance of our MolKD model, we use 9 benchmark datasets curated from MoleculeNet Wu et al. (2018) including both classification and regression tasks. These datasets cover a wide range of categories of molecular tasks, *i.e.*, quantum mechanics, physical chemistry, biophysics, and physiology. All datasets are randomly split into training, validation, and testing subsets following an 80/10/10 ratio. Furthermore, in order to investigate the robustness of the proposed model, we adopt the perturbed PhysProp dataset to estimate the PASP (property-slightly-affected structure perturbation) property Li et al. (2022b). Following Wu et al. (2018); Li et al. (2022b), different classification and regression metrics are used to fairly compare different methods: *ROC-AUC* (Area Under Curve of Receiver Operating Characteristics) for classification tasks, *RMSE* (Root-Mean-Square Error) for regression tasks except the QM8 dataset, *MAE* (Mean Absolute Error) for the QM8 dataset, and *Effect score* for the perturbed PhysProp dataset. The statistics of datasets we used in experiments are summarized in Table 4. We also provide a brief description of datasets as follows Huang et al. (2021); Li et al. (2022b):

- **QM8:** $21,786$ small molecules whose regression labels are electronic spectra and excited state energy calculated by multiple quantum mechanic methods.
- **ESOL:** $1,128$ common organic small molecules whose labels are water solubility (log solubility in mols per litre).
- **FreeSolv:** 642 small molecules whose regression labels are experimental and calculated hydration free energy in water.
- **Lipophilicity:** $4,200$ molecules whose regression labels are experimental results of octanol/water distribution coefficient (logD at pH 7.4).
- **HIV:** $41,127$ molecules whose classification labels are experimentally measured abilities to inhibit HIV replication.
- **BACE:** $1,513$ molecules whose classification labels are binary binding results for a set of inhibitors of human β-secretase 1 (BACE-1).
- **BBBP:** $2,039$ molecules with binary labels of blood-brain barrier penetration (permeability).
- **Tox21:** $7,831$ molecules with qualitative toxicity measurements on the biological target.
- **SIDER:** $1,427$ molecules which are collected from marketed drugs and adverse drug reactions (ADR) dataset.
- **ClinTox:** Qualitative data of $1,478$ drugs approved by the FDA and those that have failed clinical trials for toxicity reasons.
- **Perturbed PhysProp:** $14,176$ molecules structures and their corresponding lipophilicity properties (logP), whose principle is to determine an ideal perturbed molecule set with small perturbations that do not significantly affect the properties. In order to obtain the perturbed data, Li et al. (2022b) compare all possible molecule pairs in PhysProp (Tetko et al., 2001), calculate the fingerprint similarity of all molecules and their difference in logP, and pick out molecule pairs that meet the following two conditions: (1) the difference in the logP of the molecule pairs should be less than 0.2; (2) the molecular fingerprint similarity should be in the range of 0.3–1.0. These molecule pairs are then divided into three levels (range 0.8–1.0, 0.5–0.8, and 0.3–0.5 marked as levels 1, 2, and 3). Finally, those molecules that exist in all three levels constitute the perturbed PhysProp dataset. According to Li et al. (2022b), the evaluation metric of *effect score* is defined as follows: given a molecule set M with ground-truth properties Q and a trained predictor f, we predict the property set P by
$$P = f(M) \tag{10}$$
Similarly, Given the perturbed set M' with properties Q', we predict the property set P' by
$$P' = f(M') \tag{11}$$

Table 4: Statistics of the used datasets.

Category	Dataset	Data Type	Task Type	#Molecules	Avg. #Atoms	Avg. #Edges	Metric
Quantum Mechanics	QM8	SMILES, 3D coordinates	Regression	21,786	15.9	24.2	MAE
Physical Chemistry	ESOL	SMILES	Regression	1,128	13.3	13.7	RMSE
	FreeSolv	SMILES	Regression	642	8.7	8.4	RMSE
	Lipophilicity	SMILES	Regression	4,200	27.0	29.5	RMSE
Biophysics	HIV	SMILES	Classification	41,127	25.5	27.5	ROC-AUC
	BACE	SMILES	Classification	1,513	34.1	36.9	ROC-AUC
Physiology	BBBP	SMILES	Classification	2,039	24.1	26.0	ROC-AUC
	Tox21	SMILES	Classification	7,831	18.6	19.3	ROC-AUC
	SIDER	SMILES	Classification	1,427	33.6	35.4	ROC-AUC
	ClinTox	SMILES	Classification	1,478	26.2	27.9	ROC-AUC
Robustness	Perturbed PhysProp	SMILES	Regression	12,607	16.9	26.0	Effect score

Finally, the perturbation effect score Δ of method f is calculated by

$$\Delta = L(P, P') - L(Q, Q'), \tag{12}$$

where we use r.m.s.e as our distance function L.

A.2 Hyper-parameters

In the pre-training process on chemical reaction data, we use the following three commonly-used GNNs as the implementation of our molecular encoder: TAG (Du et al., 2017), GCN (Kipf & Welling, 2016), and GIN (Xu et al., 2019). We use the default hyper-parameters as introduced in the PyTorch-Geometric library for each GNN model. The number of propagation layers for all GNN models is 2, and the output dimension of GNNs is 2,048. The margin γ and α in Eq. 6 is set to 6 and 2 respectively. We train the model for 20 epochs with a batch size of $4,094$. We use the Adam (Kingma & Ba, 2014) optimizer with a learning rate of 1×10^{-4}, β_1 of 0.9, and β_2 of 0.999.

In the fine-tuning process on molecular property prediction data, the hidden dimension of the student GNN backbone model is 512, which is followed by two fully-connected layers to output the prediction. We carry out a grid search on the validation dataset to find the optimal temperature hyper-parameter τ in Eq. 7 and β in Eq. 9. We tune $\tau = \{0.05, 0.075, 0.1\}$ and $\alpha = \{0.2, 0.5, 0.8\}$. For example, we set $\tau = 0.05$ and $\alpha = 0.5$ on the QM8 dataset. We train the model for 400 epochs with a batch size of $1,024$. The optimization is conducted using Adam (Kingma & Ba, 2014) with a learning rate of 2×10^{-4}.

A.3 Experimental Settings

All experiments are conducted with the following settings:

- Operating system: Linux Red Hat 4.8.2-16
- CPU: Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz
- GPU: NVIDIA Tesla V100 SXM2 32GB
- Software versions: Python 3.8.10; Pytorch 1.9.0+cu102; Numpy 1.20.3; SciPy 1.7.1; Pandas 1.3.4; Scikit-learn 1.0.1; PyTorch-geometric 2.0.2; DGL 0.7.2; Open Graph Benchmark 1.3.2

B More Experimental Results

In this section, we first present the experimental results of molecular property prediction on regression tasks. Then, we investigate the choice of GNN backbone models and knowledge distillation methods as ablation studies. Finally, We provide full cases on USPTO-500-MT and PhysProp to illustrate that the proposed method can serve as an effective predictor with high robustness and interpretability.

Table 5: Results of molecular property prediction on regression tasks (metric: RMSE for ESOL and FreeSolv; MAE for QM8).

	ESOL	FreeSolv	QM8
AttentiveFP	$0.88_{\pm 0.03}$	$2.07_{\pm 0.18}$	$0.0179_{\pm 0.0001}$
GCN	$1.43_{\pm 0.05}$	$2.87_{\pm 0.14}$	$0.0366_{\pm 0.0011}$
D-MPNN	$0.98_{\pm 0.26}$	$2.18_{\pm 0.91}$	$\underline{0.0143_{\pm 0.0022}}$
SchNet	$1.05_{\pm 0.06}$	$3.22_{\pm 0.76}$	$0.0204_{\pm 0.0021}$
PretrainGNN	$1.10_{\pm 0.01}$	$2.76_{\pm 0.02}$	$0.0200_{\pm 0.0001}$
GROVER-base	$0.98_{\pm 0.09}$	$2.18_{\pm 0.05}$	$0.0218_{\pm 0.0004}$
MolCLR-GCN	$1.16_{\pm 0.00}$	$2.39_{\pm 0.14}$	$0.0181_{\pm 0.0002}$
MolCLR-GIN	$1.11_{\pm 0.01}$	$2.20_{\pm 0.20}$	$0.0174_{\pm 0.0013}$
GEM	$\underline{0.80_{\pm 0.03}}$	$1.88_{\pm 0.09}$	$0.0171_{\pm 0.0001}$
GraphMVP	1.09	-	-
MolKD^{-}	$0.93_{\pm 0.06}$	$\underline{1.64_{\pm 0.12}}$	$0.0206_{\pm 0.0028}$
MolKD (ours)	$\mathbf{0.75_{\pm 0.03}}$	$\mathbf{1.56_{\pm 0.16}}$	$\mathbf{0.0133_{\pm 0.0008}}$

Table 6: Performance on Tox21 among different student GNN models and knowledge distillation methods. The first and the second block in the *Distillation* rows are the logit-based and feature-based knowledge distillation methods, respectively.

		TAG	TAG	TAG	GCN	GIN
Hidden dim.		512	256	1024	512	512
#Params (M)		0.84	0.49	2.24	0.81	1.34
Sup.	Supervised student	$0.792_{\pm 0.054}$	$0.764_{\pm 0.083}$	$0.780_{\pm 0.049}$	$0.782_{\pm 0.052}$	$0.797_{\pm 0.047}$
Distillation	KD	$0.817_{\pm 0.042}$	$0.782_{\pm 0.042}$	$0.815_{\pm 0.063}$	$0.835_{\pm 0.055}$	$0.797_{\pm 0.040}$
	FitNet	$0.830_{\pm 0.030}$	$0.792_{\pm 0.058}$	$0.794_{\pm 0.036}$	$0.815_{\pm 0.064}$	$0.801_{\pm 0.050}$
	OFD	$0.813_{\pm 0.023}$	$0.793_{\pm 0.037}$	$0.826_{\pm 0.026}$	$0.811_{\pm 0.024}$	$0.824_{\pm 0.029}$
	MolKD	$0.841_{\pm 0.032}$	$0.802_{\pm 0.023}$	$0.848_{\pm 0.049}$	$0.837_{\pm 0.038}$	$0.831_{\pm 0.033}$

B.1 Molecular Property Prediction on Regression Tasks

We further represent the results of molecular property prediction on regression tasks. As shown in Fig. 5, MolKD consistently outperforms other competitive baseline models on three regression tasks in molecular property prediction. For example, MolKD achieves $[0.51, 0.32]$ absolute RMSE gain on FreeSolv compared with the best-performed method [AttentiveFP (w/o pre-training), GEM (with pre-training)] and $[7.52\%, 28.57\%]$ MAE gain on QM8 compared with the best-performed method [D-MPNN (w/o pre-training), GEM (with pre-training)]. These results further highlight the effectiveness of our proposed model.

B.2 Ablation Study

In MolKD, we adopt the commonly-used TAG (Du et al., 2017) as the GNN backbone models and contrastive representation distillation (Tian et al., 2019) as the knowledge distillation method. In order to show the extensibility of MolKD, we compare MolKD with other student GNN models (GCN (Kipf & Welling, 2016) and GIN (Xu et al., 2019)) and knowledge distillation methods(KD (Hinton et al., 2015), FitNet (Romero et al., 2014), and OFD (Heo et al., 2019)). For a fair comparison, we fix the pre-trained teacher GNN model, which achieves an AUC-ROC of 0.818. In Table 6, we find that the performance of MolKD is relatively stable across different student GNN models and knowledge distillation methods. The performance becomes saturated when the hidden dimension is larger than 512. We assume the cause might be that the size of molecule data is much smaller than that of reaction data, thus the number of parameters in GNN backbone models tends to be small correspondingly.

B.3 FULL CASE STUDY ON USPTO-500-MT

We select representative reactions in the test data of USPTO-500-MT as a case study in Fig. 6. We obtain the predicted product by calculating the closest molecule with respect to the sum of molecular representations of reactants learned by the corresponding algorithm. We find that the predicted product by MolKD is the same as the ground-truth in most cases (8 out of 10 cases), while Mol2vec and MolBERT fail to predict the exact product of each reaction. These results highlight the effectiveness of yield-guided chemical reaction pre-training models in MolKD. In the last two cases of Fig. 6, we can find that these two reactions are indeed very hard to predict and all three methods predict wrongly on some small atomic functional groups, such as -OH, and -COOH.

B.4 FULL INTERPRETATION CASES ON PHYSPROP

Fig. 7 presents a case study of solubility prediction with atom-level interpretation. These results are in with the intuition of chemists that the hydrophilic group (*e.g.*, -OH, and -COOH) tends to be bluer in our visualization, while the lipophilic group (*e.g.*, benzene ring) tends to be redder in our visualization. These observations indicate that our proposed MolKD model can distinguish essential atomic groups with clear interpretability of their solubility.

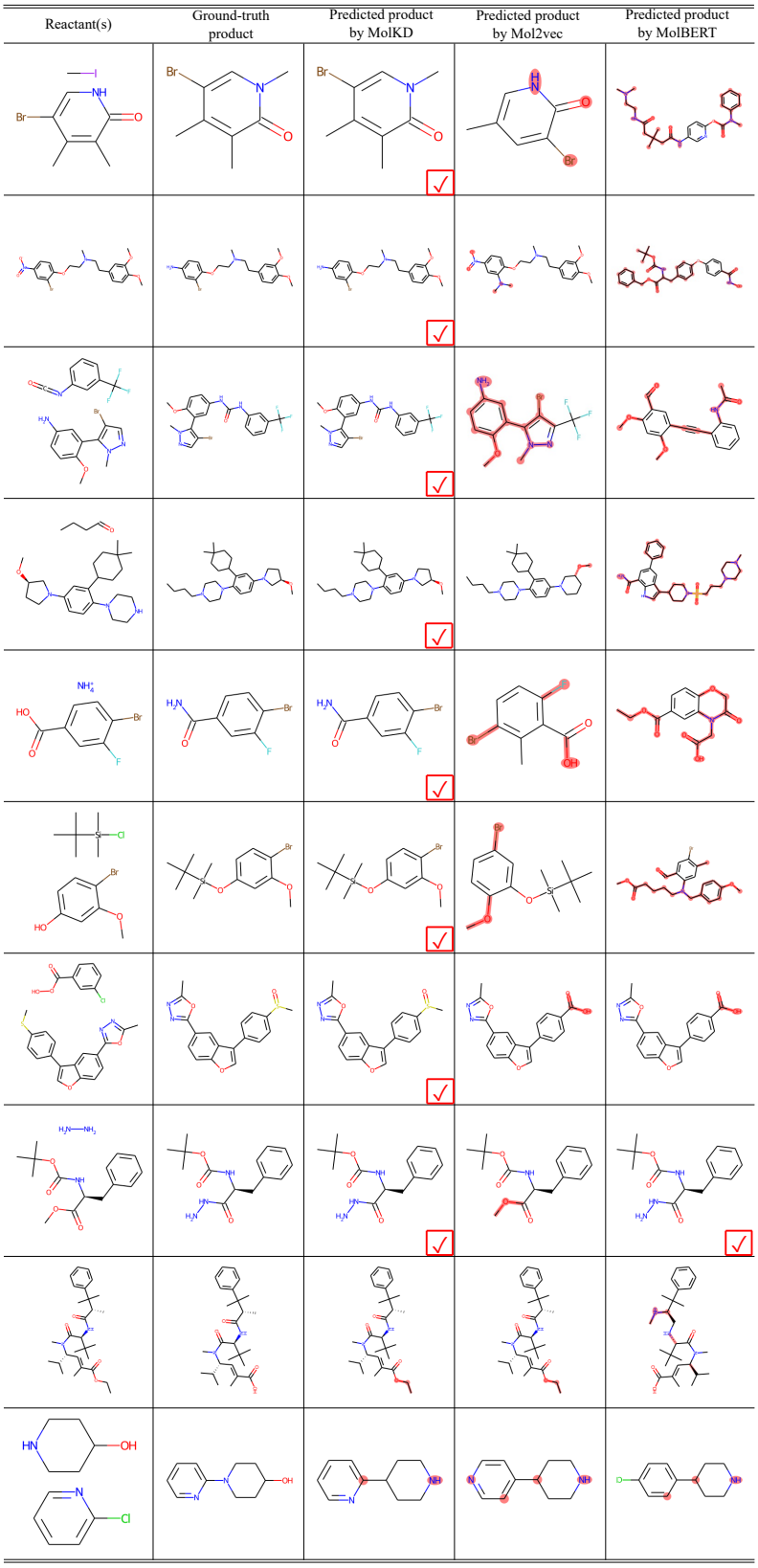

Figure 6: Case study on the USPTO-500-MT dataset. Atoms and bonds that do not match the ground-truth molecule are highlighted in red.

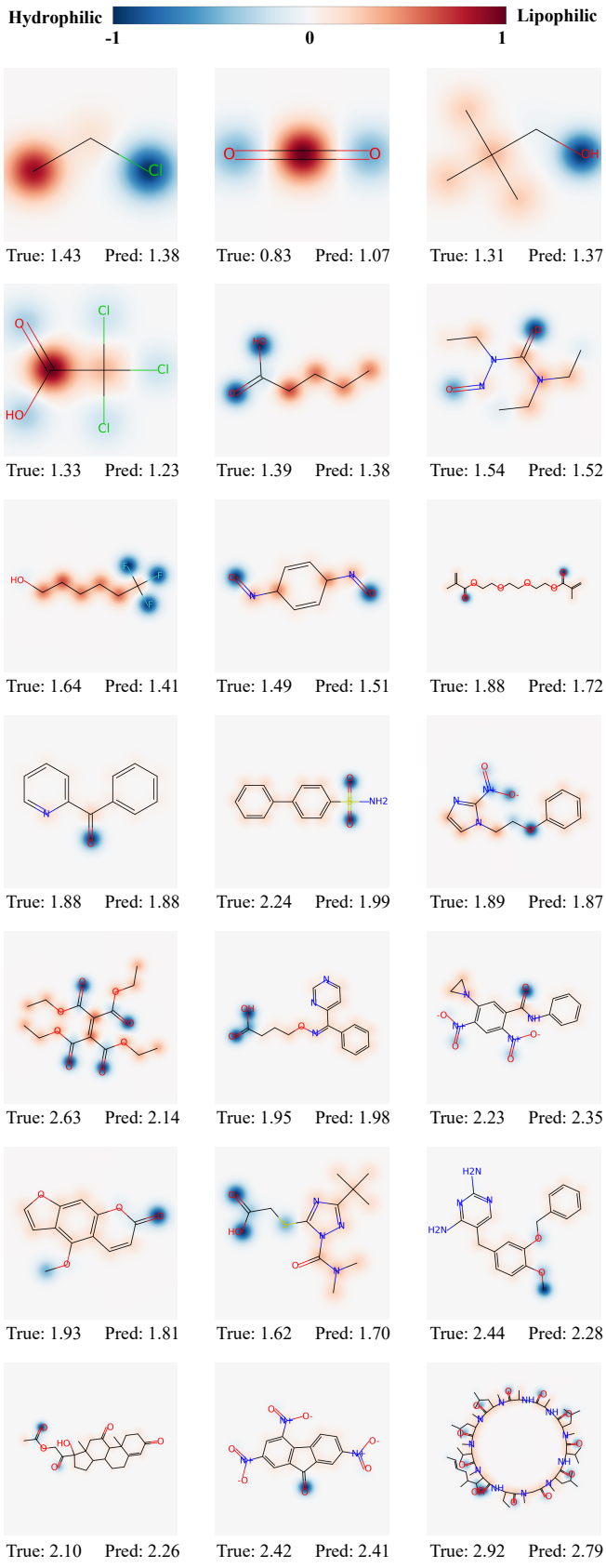

Figure 7: Case study of atom-level interpretation with true and predicted solubility labeled.