An Image-enhanced Molecular Graph Representation Learning Framework

Hongxin Xiang[1,2], **Shuting Jin**[3], **Jun Xia**[4], **Man Zhou**[5], **Jianmin Wang**[6], **Li Zeng**[2], **Xiangxiang Zeng**[1,*]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[2]Department of AIDD, Shanghai Yuyao Biotechnology Co., Ltd., Shanghai, China
[3]School of Computer Science & Technology, Wuhan University of Science and Technology, Wuhan, China
[4]School of Engineering, Westlake University, Hangzhou, China
[5]University of Science and Technology of China, Hefei, China
[6]The Interdisciplinary Graduate Program in Integrative Biotechnology, Yonsei University, Incheon, Korea

Abstract

Extracting rich molecular representation is a crucial prerequisite for accurate drug discovery. Recent molecular representation learning methods achieve impressive progress, but the paradigm of learning from a single modality gradually encounters the bottleneck of limited representation capabilities. In this work, we fully consider the rich visual information contained in 3D conformation molecular images (i.e., texture, shadow, color and planar spatial information) and distill graph-based models for more discriminative drug discovery. Specifically, we propose an image-enhanced molecular graph representation learning framework (called **IEM**) that leverages multi-view molecular images rendered from 3D conformations to boost molecular graph representations. To extract useful auxiliary knowledge from multi-view images, we design a teacher, which is pre-trained on 2 million molecules with conformations through five meticulously designed pre-training tasks. To transfer knowledge from teacher to graph-based students, we pose an efficient cross-modal knowledge distillation strategy with knowledge enhancer and task enhancer. It is worth noting that the distillation architecture of IEM can be directly integrated into existing graph-based models, and significantly improves the capabilities of these models (e.g. GIN, EdgePred, GraphMVP, MoleBERT) for molecular representation learning. In particular, GraphMVP and MoleBERT equipped with IEM achieve new state-of-the-art performance on MoleculeNet benchmark, achieving average 73.89% and 73.81% ROC-AUC, respectively. Code is available at https://github.com/HongxinXiang/IEM.

1 Introduction

The molecular representation learning plays an import role in high-precision drug discovery (such as molecular property prediction, target activity prediction) [Hu *et al.*, 2020a; Zeng *et al.*, 2022; Zhang *et al.*, 2023]. As the most direct representation of molecules, graphs treat atoms and bonds in molecule as nodes and edges in graph and have been widely studied and applied [Hu *et al.*, 2020a; Liu *et al.*, 2021; Xia *et al.*, 2023]. In view of the objective existence of molecules in nature, molecules can also be represented as images, allowing us to directly observe the morphology and internal structure of molecules through an electron microscope [Shen *et al.*, 2022] or a rendering tool [Landrum, 2013].

Limited by a single modality [Hu *et al.*, 2020a; Sun *et al.*, 2020], the GraphMVP [Liu *et al.*, 2021] and 3D InfoMax [Stärk *et al.*, 2021] try to leverage 2 modalities (2D and 3D graphs) to enhance features but the performance improvements are still limited. There are two reasons: (1) similar modalities (2D graphs and 3D graphs) and encoding ways (graph neural network-based encoder [Wu *et al.*, 2020]) and (2) weak feature extraction ability, resulting in insufficient complementary information between modalities [Tumer and Ghosh, 1995]. Image is a modality that is significantly different from graph in terms of modality and encoding way (convolutional neural network-based encoder [Li *et al.*, 2021b]). We empirically prove that 2D graph has high pearson correlation [Cohen *et al.*, 2009] with 3D graph and has significantly low pearson correlation with image, as shown above in Figure 1(a) (See Appendix A for more details[1]). Unlike graph, images understand molecules from a visual perspective, which contains texture information of molecules and allows for direct visualization of spatial arrangements without introducing any conformation, such as chiral changes of molecules. We further illustrate the advantages of images in basic prior knowledge, as shown below in Figure 1(a) (See Appendix B for more details). Given the popularity of graph neural networks in drug discovery, a meaningful question is raised: *can we exploit the rich information in molecular images to facilitate representation learning of molecular graphs?*

Considering that the simplest multi-modal learning framework [Song *et al.*, 2022; Wang *et al.*, 2020] requires additional computational costs in the training and inference stages, inspired by knowledge distillation (KD) [Hinton *et al.*, 2015; Hou *et al.*, 2021], we describe the process of information

*Correspondding author (xzeng@hnu.edu.cn)

[1]Appendix is available at https://github.com/HongxinXiang/IEM/blob/main/assets/appendix.pdf

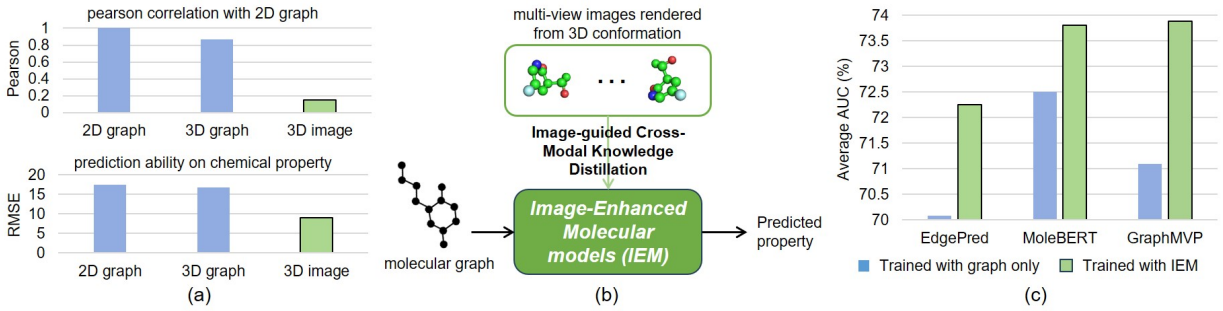

Figure 1: (a) Correlation coefficients for different molecular representation are described above. The RMSE performance of different molecular representation on chemical property $\mathcal{S}^{prop}$ is described below. (b) The proposed general cross-modal knowledge distillation framework based on multi-view molecular images (called IEM). It only introduces the prior of the image into the graph-based model during the training phase without modifying any baseline model. (c) Average ROC-AUC performance (%) on 8 classification-based property prediction tasks by using or not using the proposed IEM framework. Noticeable improvements can be observed after using the proposed IEM framework.

transfer as how to use a knowledgeable teacher (image) to teach an excellent student (graph). The success of KD depends on good teacher and KD strategies (See Section 3.6 for proof). In order to obtain a knowledgeable image-based teacher, we propose a novel multi-view molecular image representation learning method with 3D conformation to enrich the representation with 5 pre-training strategies. Considering that naive feature alignment between cross-modal features will lead to limited gain or even negative transfer [Yang *et al.*, 2021; Yan *et al.*, 2022] in the cross-modal KD (CMKD) stage, we design a novel CMKD framework (called IEM) to enrich the features of graph, which alleviates knowledge enhancer and task enhancer to align graph and image in logit space to avoid modality gaps in feature space [Liang *et al.*, 2022]. As shown in Figure 1(b), IEM distills features from images into graph representations. It is worth noting that compared with multi-modal learning methods, the proposed IEM has the following advantages: **(1) Universality:** IEM can be integrated with any graph-based method. **(2) Effectiveness:** IEM significantly improves the performance of several graph-based baselines, such as Figure 1 (c). **(3) Efficiency:** As low as 5% of training images can still improve performance; **(4) Compatibility:** IEM is compatible with both 2D and 3D molecular images and different rendering strategies.

In summary, the main contributions of this paper are summaried as follows:

- To the best of our knowledge, we are the first to reconstruct the graph-based molecular representation learning into image-based cross-modal distillation paradigm.
- We propose the 3D conformational molecular image representation method to train a knowledgeable teacher, achieving a one-time pre-training on 2 million molecules that enables the use of multi-view features to enhance existing graph-based baseline models.
- We propose an image-enhanced molecular graph representation learning framework, called IEM, which equips with knowledge enhancer and task enhancer to improve the performance of a large number of graph-based models. In addition, we also demonstrate the effectiveness of our framework theoretically.
- We show that our method achieves significantly better performance on 12 MoleculeNet benchmarks and can substantially enhance the performance of existing molecular graph representation models.

2 Related Work

Graph-based Molecular Representation Learning. Graph neural networks have achieved remarkable success in drug discovery tasks. In view of the high cost of annotating molecules, recent studies mainly learn from large-scale label-free molecular databases by designing pre-training strategies. [Hu *et al.*, 2020a; Li *et al.*, 2021a; Guo *et al.*, 2021] consider both node-level and graph-level pre-training strategies to capture the local and the global information in graph. GPT-GNN [Hu *et al.*, 2020b] uses two generative pre-training tasks, including attribute generation and edge generation, to extract fine-grained information of molecules. GROVER [Rong *et al.*, 2020] and MGSSL [Zhang *et al.*, 2021] propose a motif-based prediction or generation tasks to capture the information of molecular motifs. Mole-BERT [Xia *et al.*, 2023] proposed masked atoms modeling and triplet masked contrastive learning to further optimize mask-based GNN. Considering the importance of 3D geometric information in drug discovery tasks, GraphMVP [Rong *et al.*, 2020] and 3D Infomax [Stärk *et al.*, 2021] pre-trained GNN on molecules with 3D geometric information. In the paper, we still follow the graph-based paradigm for drug discovery tasks and try to improve all graph-based models.

Image-based Molecular Representation Learning. Because graphs are discrete and unordered, some researchers consider representing molecules as images and utilizing mature computer vision techniques to extract features. Chemception [Goh *et al.*, 2017], 2DConvNet [Fernandez *et al.*, 2018] and DenseNet121 [Zhong *et al.*, 2021] use molecular images to predict chemical properties, toxicity of compounds and contaminant reactivity, respectively. With the explosion of self-supervised learning in computer vision [Arnab *et al.*, 2021; Likhosherstov *et al.*, 2021], the first pre-training model based on molecular images (called ImageMol) is proposed for learning representation from 10 million molecules with 5 well-designed pre-training tasks [Zeng *et al.*, 2022]. CGIP is further proposed to extract fine-grained image representations via carefully designed intra- and inter-modal contrastive learning between graph and image [Xiang *et al.*, 2023]. Different from

the images used by previous methods, which are 2D images generated by RDKit, we propose a clearer and more informative 3D conformation-based multi-view image representation.

Cross-Modal Knowledge Distillation. As an important branch of knowledge distillation [Hinton *et al.*, 2015], cross-modal knowledge distillation (CMKD) is still a relatively emerging field, which refers to using a teacher from another modality to supervise the learning model of the current modality and improve the performance of the student during inference. For example, [Gupta *et al.*, 2016] transfer supervision from labeled RGB images to unlabeled depth and optical flow images and [Sun *et al.*, 2021] learns TIR (Thermal Infrared)-specific target representations transferred from the RGB modality. Pri3D [Hou *et al.*, 2021] distills 3D point cloud information into 2D images and improves performance of image encoder in semantic segmentation, object detection, and instance segmentation. PointCMT [Yan *et al.*, 2022] and UniDistill [Zhou *et al.*, 2023] utilize images as teachers to guide point cloud-based students and airbornelidar-based students for improving point cloud-related tasks and 3D object detection in bird's-eye view, respectively. VGSR [Jin *et al.*, 2023] distills the knowledge of face image into audio to improve the performance of speaker recognition. Different from previous works, this is the first image-to-graph cross-modal knowledge distillation framework to our best knowledge.

3 Our Method

3.1 Preliminaries

Background

We summarized three ways to obtain molecular images: (1) Canvas-based technology, which draws molecular images by creating a sketchpad and using pixels, such as RDKit [Landrum, 2013]; (2) 3D CAD (Computer Aided Design) modeling technology, which uses CAD software to create a virtual three-dimensional space and build a model with three-dimensional data or geometric configuration, such as PyMol [DeLano and others, 2002]; (3) Physical microscopy-based technology, including Cryo-EM and single-molecule fluorescence imaging, which utilizes atomic force microscopy (AFM), optical tweezers (OT), magnetic tweezers (MT) or fluorescence microscopy to obtain visual representations of molecules. See Appendix C for the visualization of molecules in three ways. Each of these methods has its own advantages. Canvas technology has low computational complexity and fast processing speed, but is confusing when describing complex, geometric molecules. Existing molecular image-based methods are all based on this imaging method [Zeng *et al.*, 2022; Xiang *et al.*, 2023]. 3D CAD modeling technology retains rich information in molecules but is slow to render. Single-molecule fluorescence imaging technology can directly reflect the molecular portraits of nature, but obtaining these samples is low-throughput and resource-consuming, making it impossible to expand to large-scale data. Therefore, in this paper, we only consider the first two imaging methods. Moreover, the proposed method can seamlessly support samples collected by the third technology.

Problem Formulation

Let the molecular graphs and corresponding ground-truth labels on downstream tasks are $\{\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)\}_{i=1}^n$ and $\{y_i\}_{i=1}^n \in \mathbb{R}^t$ respectively, where $\mathcal{V}_i \in \mathbb{R}^{n_i^v \times d_i^v}$ with the number of vertices is n_i^v and the feature dimension of the vertices is d_i^v and $\mathcal{E}_i \in \mathbb{R}^{n_i^e \times d_i^e}$ with the number of edges is n_i^e and the feature dimension of the edges is d_i^e represent vertices and edges in i-th molecular graph respectively. t is the number of tasks in downstream tasks. The corresponding single-view 2D images and multi-view 3D images of molecular graphs $\mathcal{G}$ can be denoted as $\mathcal{V}^{2D} \in \mathbb{R}^{H \times W \times 3}$ and $\mathcal{V}^{3D} \in \mathbb{R}^{V \times H \times W \times 3}$ respectively, where V represents the number of views and H and W represent the height and the width of images, respectively. The single-view 2D images can be directly generated by RDKit [Landrum, 2013] and the multi-view 3D images can be obtained by using PyMol [DeLano and others, 2002] to generate snapshots from different viewpoints. The features of graphs $\mathcal{F}^g \in \mathbb{R}^{d^g}$, 2D images $\mathcal{F}^{2D} \in \mathbb{R}^{d^v}$ and 3D images $\mathcal{F}^{3D} \in \mathbb{R}^{d^v}$ can be extracted by graph encoder Enc^g, 2D encoder Enc^{2D} and 3D encoder Enc^{3D}, respectively. In the proposed knowledge distillation framework, after pre-training, both Enc^{2D} and Enc^{3D} from images can be chosen as teacher, while graph encoder is considered as student. This ensures that our proposed knowledge distillation framework remains applicable when it is necessary to perform representation learning on molecular graphs for which 3D conformation cannot be acquired. Considering that 3D will carry more feature information, our experiments choose the Enc^{3D} as the teacher.

3.2 Overview of the Method

Here, we propose the image-enhanced molecular graph representation learning framework (IEM), which equips knowledgeable teachers and distillation strategies to prevent negative transfer. In particular, we pre-train the teacher with 5 pre-tasks to ensure that the features extracted by the teacher are better than those of the students, where 1 pre-task takes into account the intrinsic information of molecular images by aligning 2D and 3D images and 4 pre-tasks consider four priors of molecule (atom $\mathcal{S}^{atom} \in \mathbb{R}^{n^{atom}}$, bound $\mathcal{S}^{bound} \in \mathbb{R}^{n^{bound}}$, geometry $\mathcal{S}^{geom} \in \mathbb{R}^{n^{geom}}$ and basic properties $\mathcal{S}^{prop} \in \mathbb{R}^{n^{prop}}$) by using 4 corresponding predictors (P^{atom}, P^{bound}, P^{geom} and P^{prop}) to optimize molecular features. These four predictors use the same structure: Full Connection (FC)→Softplus→FC. During the distillation stage, simply aligning features from different modalities dimension-wise can easily lead to negative transfer [Yang *et al.*, 2021; Yan *et al.*, 2022]. Inspired by them, we utilize knowledge enhancer Enh^k and task enhancer Enh^t to align the information of different modalities at the logit level, which preserves both modality-specific semantics from student and logit-specific semantics from teacher guidance. The overview of IEM is illustrated in Figure 2(a). We summarize the main processes in Appendix D. The process of IEM is divided into three steps: (1) Use 5 pre-training tasks to train a knowledgeable teacher (Section 3.3); (2) Exploit image-based teacher to enhance graph-based student by using the knowledge enhancer and task enhancer (Section 3.4); (3) Train IEM and inference in downstream tasks (Section 3.5).

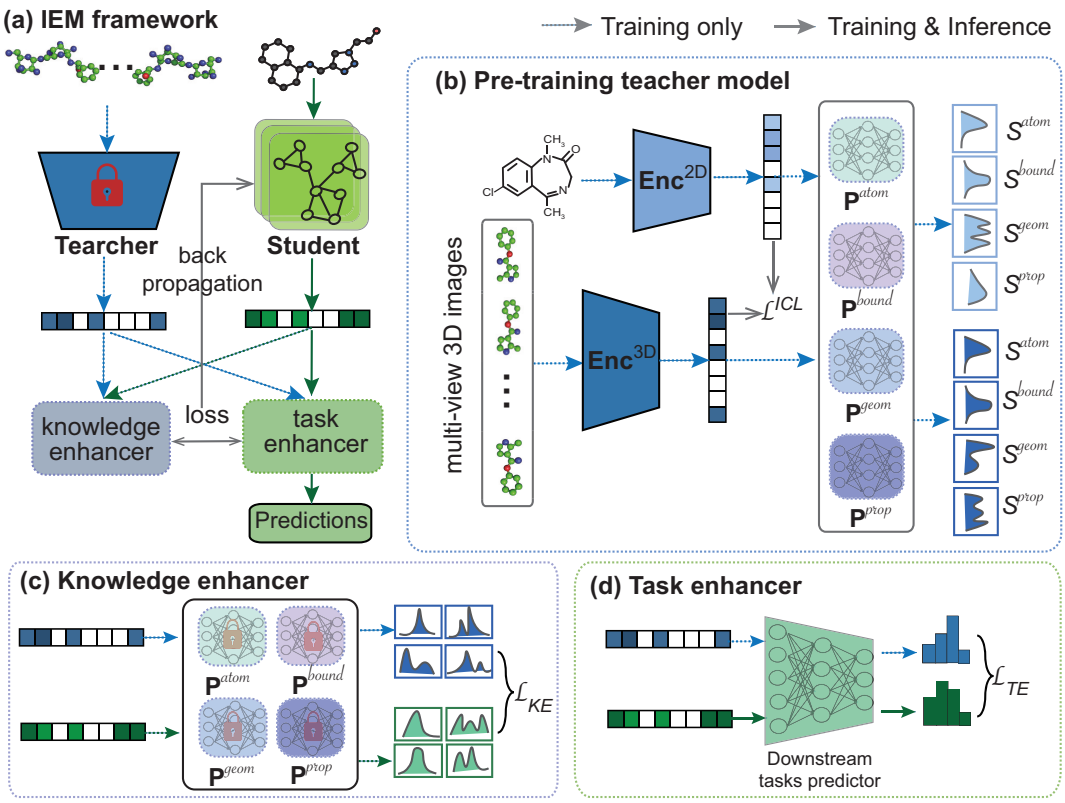

Figure 2: Overview of the proposed IEM framework. **(a)** Multi-view molecular images rendered from 3D conformations are forward propagated into the frozen teacher, which is the 3D image encoder pretrained in (b), to extract visual features. Knowledge enhancer and task enhancer distill knowledge from visual features into graph-based models. The purpose of the knowledge enhancer is to enhance features relevant to the upstream/pre-trained tasks by aligning the output distributions from the 4 predictors in (b). The purpose of the task enhancer is to enhance information related to downstream tasks by aligning the output distribution of the task predictor. **(b)** The process of pre-training the teacher. **(c)** Execution process of the knowledge enhancer. **(d)** Execution process of the task enhancer.

3.3 Pretraining Teacher

Excellent teacher are one of the keys to knowledge distillation. In this paper, as shown in Figure 2(b), we combine low-computation cost 2D images generated by RDKit and information-rich 3D images rendered by PyMol, both of which can be used to enhance graph representation (See Appendix E for details of image rendering). However, we still recommend using a 3D image encoder as a teacher because it is more discriminative. Here, we choose 4 fixed viewpoints to generate multi-view images. The feature extraction process can be formalized as: $\mathcal{F}^{2D} = \text{Enc}^{2D}(\mathcal{V}^{2D})$ and $\mathcal{F}^{3D} = \text{Enc}^{3D}(\mathcal{V}^{3D})$. Due to the uncertainty of downstream tasks, the teacher should have strong knowledge generalization capabilities. Therefore, we designed 5 pre-training tasks based on the principle that generalized knowledge should reflect the basic attribute information of molecules, including (1) contrastive learning between 2D and 3D images (ICL); (2) atom distribution prediction task (ADP); (3) bound distribution prediction task (BDP); (4) geometry distribution prediction task (GDP); (5) property distribution prediction task (PDP).

The main motivation of the ICL is to extract the inherent information in molecular images. Meanwhile, the 2D image encoder has a perception of 3D information and the 3D image encoder has a perception of global information to alleviate the problem of view-occlusion. For convenience, we denote $\mathcal{F}^I = \{\mathcal{F}^{2D}, \mathcal{F}^{3D}\} \in \mathbb{R}^{2\times n, d^v}$, where n and d^v represent the number of samples and the dimension of visual features. Following InfoNCE [Oord *et al.*, 2018], the contrastive learning loss between 2D and 3D images is formalized as:

$$\mathcal{L}_{ICL} = -\frac{1}{2n}\sum_{i=1}^{2n} log\frac{\exp(sim(\mathcal{F}_i^I, \mathcal{F}_{2n+i}^I)/\tau)}{\sum_{j=1}^{2n}\mathbb{I}_{i\neq j}\exp(sim(\mathcal{F}_i^I, \mathcal{F}_j^I)/\tau)} \quad (1)$$

where $\mathbb{I}$ and τ represent indicator function and temperature, respectively. $sim(\cdot)$ represents cosine similarity, that is, $\text{sim}(\mathcal{F}_i^I, \mathcal{F}_j^I) = \mathcal{F}_i^I \cdot \mathcal{F}_j^I / \parallel \mathcal{F}_i^I \parallel_2 \parallel \mathcal{F}_j^I \parallel_2$.

The motivation for the ADP, BDP, GDP and PDP tasks is to allow the teacher to acquire knowledge with strong generalization. Therefore, we extract fundamental prior knowledge of molecules from their atoms (S^{atom}), bonds (S^{bound}), geometry (S^{geom}), and chemical properties (S^{prop}). For details on these priors, see Appendix F. Subsequently, $\mathcal{F}^{2D}$ and $\mathcal{F}^{3D}$ are forward-propagated to 4 predictors $\text{P}^{atom}, \text{P}^{bound}, \text{P}^{geom}, \text{P}^{prop}$ to obtain the corresponding predicted logits $\{p^{2D}_{atom}, p^{2D}_{bound}, p^{2D}_{geom}, p^{2D}_{prop}\}$ and $\{p^{3D}_{atom}, p^{3D}_{bound}, p^{3D}_{geom}, p^{3D}_{prop}\}$. Finally, the losses of 2D im-

age encoder and 3D image encoder in these 4 prediction tasks can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{2D} = &\left|p_{atom}^{2D} - \mathcal{S}^{atom}\right| + \left|p_{bound}^{2D} - \mathcal{S}^{bound}\right| \\ &+ \left|p_{geom}^{2D} - \mathcal{S}^{geom}\right| + \left|p_{prop}^{2D} - \mathcal{S}^{prop}\right|\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{3D} = &\left|p_{atom}^{3D} - \mathcal{S}^{atom}\right| + \left|p_{bound}^{3D} - \mathcal{S}^{bound}\right| \\ &+ \left|p_{geom}^{3D} - \mathcal{S}^{geom}\right| + \left|p_{prop}^{3D} - \mathcal{S}^{prop}\right|\end{aligned} \tag{3}$$

Finally, the 2D image encoder and 3D image encoder can be pretrained with the following total loss:

$$\mathcal{L}_{Pretrain}^{Teacher} = \mathcal{L}_{ICL} + \mathcal{L}_{2D} + \mathcal{L}_{3D} \tag{4}$$

See Appendix G for mode details about pretraining teacher.

3.4 Image-enhanced Distillation Strategy

During CMKD, both 2D image encoder and 3D image encoder can be considered teachers. Here, we take a 3D image encoder as a teacher as an example. We first use frozen teacher and untrained GNN student to extract the corresponding features $\mathcal{F}^v \in \mathbb{R}^{d^v}$ and $\mathcal{F}^g \in \mathbb{R}^{d^g}$. Considering that the output dimensions of the teacher and the student may be different, we add a fully connected layer after the student network by default to make $d^g = d^v$. Subsequently, instead of using simple feature alignment we use knowledge enhancer Enh^k and task enhancer Enh^t to transfer knowledge from $\mathcal{F}^{3D}$ to $\mathcal{F}^g$, which can alleviate the problem of negative transfer.

Knowledge Enhancer: The purpose of the knowledge enhancer is to distill task-irrelevant generalized knowledge from the teacher. As shown in Figure 2(c), Enh^k accepts $\mathcal{F}^{3D}$ and $\mathcal{F}^g$ as input and uses 4 frozen predictors $\text{P}^{atom}, \text{P}^{bound}, \text{P}^{geom}, \text{P}^{prop}$ to predict the distribution of molecules at different knowledge levels, which can be formalized as $\{p_{atom}^v, p_{bound}^v, p_{geom}^v, p_{prop}^v\}$ and $\{p_{atom}^g, p_{bound}^g, p_{geom}^g, p_{prop}^g\}$. We use the predicted labels of the teacher as ground-truth to supervise students with L1 loss, which can be formalized as:

$$\begin{aligned}\mathcal{L}_{KE} = &\left|p_{atom}^g - p_{atom}^v\right| + \left|p_{bound}^g - p_{bound}^v\right| \\ &+ \left|p_{geom}^g - p_{geom}^v\right| + \left|p_{prop}^g - p_{prop}^v\right|\end{aligned} \tag{5}$$

It is worth noting that this paper does not use KL-based distillation loss [Hinton *et al.*, 2015] because it is difficult to apply to single-task classification and regression tasks.

Task Enhancer: The purpose of the task enhancer is to distill knowledge relevant to downstream tasks from teacher. As shown in Figure 2(d), task predictor accepts $\mathcal{F}^{3D}$ and $\mathcal{F}^g$ and generates task-related prediction logits $\bar{y}^v$ and $\bar{y}^g$. We use $\bar{y}^v$ from the teacher as ground-truth to supervise $\bar{y}^g$ from the student with smooth L1 loss [Girshick, 2015], which can be formalized as:

$$\mathcal{L}_{TE} = \begin{cases} 0.5(|\bar{y}^g - \bar{y}^v|)^2 \text{ if } |\bar{y}^g - \bar{y}^v| < 1 \\ |\bar{y}^g - \bar{y}^v| - 0.5 \text{ otherwise} \end{cases} \tag{6}$$

3.5 Training and Inference

For efficiency, we usually hope to use a single modality rather than multiple modalities when inferring. We further use the true ground-truth y of the downstream task to train student, and the loss is formalized as:

$$\mathcal{L}_T = \varphi(\bar{y}^g, y) \tag{7}$$

where φ represents cross-entropy loss for classification and MSE loss for regression. The final loss is formalized as:

$$\mathcal{L}_{total} = \lambda_{KE}\mathcal{L}_{KE} + \lambda_{TE}\mathcal{L}_{TE} + \mathcal{L}_T \tag{8}$$

During inference, predicted results can be obtained by inputting graph data $\mathcal{G}$ to the student network Enc^g.

3.6 Theoretical Justification of IEM Effectiveness

We utilize CMKD to transfer knowledge from image-based teachers to graph-based students to obtain image-enhanced graph features $\mathcal{F}^{IE}$. For a useful knowledge distillation, F^{IE} with knowledge γ from teacher should be more informative than F^g from student only. We denote the information enhancement useful to the task after knowledge distillation as $I_{diff} = \mathcal{I}^{IE} - \mathcal{I}^g$, where $\mathcal{I}^{IE} = \mathcal{I}(\mathcal{F}^{IE}|\mathcal{V}, y; \text{Enc}^g, \gamma)$ and $\mathcal{I}^g = \mathcal{I}(\mathcal{F}^g|\mathcal{G}, y; \text{Enc}^g)$ represent the information amount of features enhanced by teacher and features extracted only by student, respectively. $\mathcal{I}(\bullet|\star)$ represents the amount of information produced by $\bullet$ under the conditions of given $\star$. We prove that when negative transfer does not occur, the lower bound Ω of the information increment $\mathcal{I}_{diff}$ depends on the information difference between the teacher and the student, which can be formulated as $\Omega = \mathcal{I}(\mathcal{F}^{3D}|\mathcal{V}, y; \text{Enc}^{3D})$-$\mathcal{I}(\mathcal{F}^g|\mathcal{G}, y; \text{Enc}^g)$. Therefore, a knowledgeable teacher and a effective distillation strategy to prevent negative transfer are crucial for CMKD. Please see Appendix H for detailed proof. To ensure effectiveness of IEM, we design teacher with 5 pre-training tasks and 2 enhancers for alleviating negative transfer.

4 Experiments and Results

4.1 Experimental Settings

Datasets and evaluation protocol. For pre-training teacher model, we sample 2 millions unlabeled molecules with 3D conformations from PCQM4Mv2 database [Hu *et al.*, 2017]. In evaluation stage, we use the widely-used 8 binary classification datasets from MoleculeNet [Wu *et al.*, 2018] with ROC-AUC metric. For broader evaluation, we also test the effectiveness of IEM on 4 regression datasets included in GraphMVP [Liu *et al.*, 2021] with root mean square error (RMSE) metric. See Appendix I for more dataset details. Notably, we use strict scaffold splitting [Hu *et al.*, 2020a] to divide all datasets into training set, validation set and test set according to 8:1:1.

Implementation details. In pre-training teacher model stage, we use 2 independent ResNet-18 [He *et al.*, 2016] as the architecture for 2D and 3D image encoders. The 2D image encoder extracts 512-dimensional visual features directly from a single 2D molecule image, while the 3D image encoder obtains 512-dimensional visual features by applying view-wise mean-pooling to the features of multi-view 3D images. In addition, the atom, binding, geometry, property predictors are multi-layer perceptrons (MLPs), including a 512-dimensional input layer, a 256-dimensional hidden layer, a softplus activation function and an output layer related to prediction. The teacher model is pre-trained for more than 30 epochs (about

	Tox21	ToxCast	Sider	ClinTox	MUV	HIV	BBBP	BACE	Average
#Molecules	7831	8576	1427	1478	93087	41127	2039	1513	-
#Task	12	617	27	2	17	1	1	1	-
GIN [Xu *et al.*, 2018]	74.3(0.9)	61.5(0.8)	57.3(1.2)	57.2(4.1)	71.6(2.8)	75.2(2.0)	66.7(1.8)	69.6(5.5)	66.68
IEM-GIN	74.5(0.4)	62.5(0.8)	59.1(1.7)	62.6(4.1)	77.7(2.9)	77.9(1.3)	69.3(1.9)	77.7(3.5)	70.16
Δ	↑ 0.2	↑ 1.0	↑ 1.8	↑ 5.4	↑ 6.1	↑ 2.7	↑ 2.6	↑ 8.1	↑ 3.5
EdgePred [Hu *et al.*, 2020a]	76.0(0.6)	64.1(0.6)	60.4(0.7)	64.1(3.7)	75.1(1.2)	76.3(1.0)	67.3(2.4)	77.3(3.5)	70.08
IEM-EdgePred	76.3(0.6)	64.6(0.6)	61.2(0.6)	67.5(2.3)	78.3(1.3)	78.3(1.3)	67.8(2.2)	**84.1(0.8)**	72.26
Δ	↑ 0.3	↑ 0.5	↑ 0.8	↑ 3.4	↑ 3.2	↑ 2.0	↑ 0.5	↑ 6.8	↑ 2.2
GraphMVP [Liu *et al.*, 2021]	74.5(0.7)	63.4(0.5)	60.7(1.4)	78.4(6.4)	73.0(2.3)	75.6(1.6)	67.4(2.4)	75.8(3.0)	71.10
IEM-GraphMVP	75.9(0.7)	64.4(0.6)	61.9(1.7)	**80.8(3.1)**	77.3(1.2)	**78.8(1.1)**	68.7(1.0)	83.3(1.4)	**73.89**
Δ	↑ 1.4	↑ 1.0	↑ 1.2	↑ 2.4	↑ 4.3	↑ 3.2	↑ 1.3	↑ 7.5	↑ 2.8
GraphMVP-C [Liu *et al.*, 2021]	74.6(0.4)	63.4(0.6)	60.6(1.3)	76.9(3.7)	72.8(2.4)	77.1(2.1)	69.9(1.4)	79.6(1.7)	71.86
IEM-GraphMVP-C	75.6(0.6)	64.8(0.5)	62.0(0.9)	79.2(2.9)	77.0(1.7)	78.2(1.0)	**71.4(1.4)**	81.9(1.6)	73.76
Δ	↑ 1.0	↑ 1.4	↑ 1.4	↑ 2.3	↑ 4.2	↑ 1.1	↑ 1.5	↑ 2.3	↑ 1.9
Mole-BERT [Xia *et al.*, 2023]	77.0(0.3)	64.4(0.2)	63.2(0.7)	72.7(2.7)	79.2(2.0)	77.7(0.7)	65.7(2.3)	80.2(0.9)	72.51
IEM-Mole-BERT	**77.8(0.4)**	**65.6(0.3)**	**65.3(0.8)**	72.2(1.4)	**79.7(1.8)**	**78.8(0.6)**	68.1(1.0)	83.0(0.9)	73.81
Δ	↑ 0.8	↑ 1.2	↑ 2.1	-0.5	↑ 0.5	↑ 1.1	↑ 2.4	↑ 2.8	↑ 1.3

Table 1: The ROC-AUC (%) performance of different methods on 8 classification datasets of molecular property prediction. We report the mean (standard deviation) ROC-AUC of 10 random seeds from 0 to 9 with scaffold splitting. The best and second best results are marked **bold** and underlined. IEM-baseline represents baseline equipped with IEM. Δ represents the absolute improvement percentage calculated by $\text{AUC}_{\text{w/ IEM}} - \text{AUC}_{\text{w/o IEM}}$.

450k steps) with temperature of 0.1, batch-size of 128 and learning rate of 0.01 (see Appendix J for details of training losses). Note that the pre-trained teacher model is able to distill any graph-based model. In order to align the experimental settings and fair comparison, we uniformly use those methods with 5-layer Graph Isomorphism Networks (GIN) [Xu *et al.*, 2018] and 300 hidden layer dimensions as our baseline models for evaluation on downstream tasks. Following [Hu *et al.*, 2020a], we train for 100 epochs with batch-size of 32 and learning rate of 0.001. We select hyper-parameters λ_{KE} and λ_{TE} from $\{0.001, 0.01, 0.1, 1, 5\}$ and report test scores corresponding to the best validation performance. Notice that the results of some baselines may differ from their original papers because inconsistent evaluation settings and we reproduced them with the same evaluation.

4.2 Main Results

We first evaluate the performance of IEM on the 8 molecular property prediction datasets with 5 baselines (GIN, EdgePred, GraphMVP, GraphMVP-C and Mole-BERT) and Table 1 reports the main results (See Appendix K for more comparison methods). We observe that the baselines equipped with IEM achieve the state-of-the-art performance. It is worth noting that all baselines obtain consistent performance improvement after being equipped with IEM with an absolute improvement ranging from 1.3% to 3.5% on average ROC-AUC. To verify that the performance improvement of IEM is not caused by the standard deviation of the baseline, we counted the results that are higher than the $\alpha = \textit{mean} + \textit{standard deviation}$ of the baseline. We find that 70% (28 out of 40) results outperform the α of the baseline. Especially for GIN, except for the Tox21, other improvements are better than the α of the baseline.

We further evaluate on a wider range of drug discovery tasks, which includes 4 regression benchmarks. As shown in Table 2, we find the same conclusion as the classification task, that is, IEM comprehensively improved all baselines with a maximum 8.56% relative improvement and achieved the state-of-the-art performance. Therefore, IEM is promising as a universal plug-in to improve any graph-based model.

	ESOL	Lipo	Malaria	CEP
#Molecules	1,128	4,200	9,999	29,978
#Task	1	1	1	1
GIN	1.472(0.038)	0.832(0.025)	1.113(0.011)	1.340(0.018)
IEM-GIN	1.346(0.045)	0.817(0.019)	1.084(0.003)	1.329(0.021)
Δ	↑ 8.56%	↑ 1.80%	↑ 2.61%	↑ 0.82%
EdgePred	1.367(0.041)	0.778(0.013)	1.110(0.011)	1.362(0.025)
IEM-EdgePred	1.350(0.027)	0.769(0.006)	1.088(0.005)	1.345(0.016)
Δ	↑ 1.24%	↑ 1.16%	↑ 1.98%	↑ 1.25%
GraphMVP	1.322(0.062)	0.773(0.016)	1.128(0.019)	1.308(0.024)
IEM-GraphMVP	1.281(0.044)	0.754(0.015)	1.089(0.005)	**1.294**(0.020)
Δ	↑ 3.10%	↑ 2.46%	↑ 3.46%	↑ 1.07%
GraphMVP-C	1.333(0.055)	0.768(0.013)	1.114(0.008)	1.304(0.020)
IEM-GraphMVP-C	1.274(0.037)	0.761(0.017)	1.090(0.004)	1.296(0.012)
Δ	↑ 4.43%	↑ 0.91%	↑ 2.15%	↑ 0.61%
MoleBERT	1.115(0.017)	0.727(0.006)	1.137(0.021)	1.350(0.015)
IEM-MoleBERT	1.090(0.031)	**0.716**(0.003)	**1.080**(0.003)	1.343(0.013)
Δ	↑ 2.24%	↑ 1.51%	↑ 5.01%	↑ 0.52%
GraphMVP-F	1.094(0.037)	0.724(0.009)	1.106(0.013)	1.397(0.040)
IEM-GraphMVP-F	**1.067**(0.039)	**0.716**(0.010)	1.093(0.012)	1.392(0.026)
Δ	↑ 2.47%	↑ 1.10%	↑ 1.18%	↑ 0.36%

Table 2: The RMSE performance on 4 regression datasets of molecular property prediction. We report the mean (standard deviation) RMSE of 10 random seeds from 0 to 9 with scaffold splitting. IEM-baseline represents baseline equipped with IEM. Δ represents the relative improvement percentage calculated by $(1 - \frac{\text{w/o IEM}}{\text{w/ IEM}}) \times 100$.

4.3 Different GNN Architectures

To verify that IEM is effective on different GNN architectures, we used 4 different GNN architectures, including GCN, GIN, GAT, and GraphSAGE. The Table 3 shows the average ROC-AUC results of different GNN architectures on 8 classification datasets. We find that IEM can significantly improve different GNN architectures with a relative performance improvement of 3.92% to 5.23%. In particular, we observe that GIN equipped with IEM is competitive with pre-trained models such as InfoGraph, EdgePred, and 3D InfoMax. This shows the strong generalization ability of the IEM, which is able to achieve performance comparable to pre-training even without any pre-training.

	GCN	GIN	GAT	GraphSAGE
w/o IEM	66.88	66.68	66.53	66.99
w/ IEM	69.81	70.16	69.76	69.61
Δ	$\uparrow$ 4.39%	$\uparrow$ 5.23%	$\uparrow$ 4.87%	$\uparrow$ 3.92%

Table 3: The average ROC-AUC (%) performance on 8 classification datasets with different GNN architectures. w/o means baseline without IEM and w/ means baseline with IEM. Δ represents the relative improvement percentage calculated by $(1 - \frac{\text{w/o IEM}}{\text{w/ IEM}}) \times 100$.

4.4 Different Image Rendering Strategies

To explore the performance of the IEM on conformation-free molecules, we discuss the impact of different image rendering strategies, including conformation-free 2D images rendered by RDKit and PyMol. As shown in Table 4. We find that IEM can improve the performance of EdgePred and GraphMVP with an average performance improvement of more than 2% in all rendering strategies, indicating that IEM can successfully be compatible with conformation-free 2D images. We also find that the performance of IEM on 2D images rendered by RDKit and PyMol is similar (72.21% v.s. 72.00% and 73.34% v.s. 73.41%), indicating that we can use the more economical RDKit to improve graph-based models under limited computing resources.

Image rendering		Method	
Image type	Rendering strategy	EdgePred	GraphMVP
×	×	70.08	71.1
2D	RDKit	72.21 ($\uparrow$ 3.04%)	73.34 ($\uparrow$ 3.15%)
2D	PyMol	72.00 ($\uparrow$ 2.74%)	73.41 ($\uparrow$ 3.25%)
3D	PyMol	**72.26** ($\uparrow$ 3.11%)	**73.89** ($\uparrow$ 3.92%)

Table 4: The average ROC-AUC (%) performance on 8 classification datasets with different image rendering methods. The number in bracket indicates the percentage of absolute performance improvement compared to the baseline without IEM.

4.5 Image Efficiency

To verify demonstrating the image-efficiency of IEM, we use different numbers of images to distill GraphMVP. Table 5 shows the average ROC-AUC with different number of images. We find that as the number of image samples increases, the performance of the IEM continues to improve with a range from 1.55% to 3.92%. In particular, using only 5% of the number of images, IEM is still able to improve the performance of baseline with a relative improvement of 1.55%, proving that IEM is efficient for images.

	image size					
	0%	5%	10%	20%	50%	100%
IEM	71.10	72.20	72.26	72.95	73.38	73.89
Δ	-	$\uparrow$ 1.55%	$\uparrow$ 1.64%	$\uparrow$ 2.60%	$\uparrow$ 3.20%	$\uparrow$ 3.92%

Table 5: The average ROC-AUC (%) performance on 8 classification datasets with different number of images. The image size represents the proportion of image samples used. We use GraphMVP as baseline model. Δ represents the relative improvement percentage.

4.6 Ablation Study

Table 6 shows the ablation results on knowledge enhancer and task enhancer using 4 different GNNs. We find that a single enhancer can improve the performance of baselines with an absolute improvement of up to 2.06%, which shows the effectiveness of two enhancers. By combining the two enhancers, more performance improvements can be achieved, with absolute performance improvements ranging from 2.62% to 3.48%. We also note that the performance improvement of KE is always better than that of TE, indicating that the fundamental prior knowledge (atom, bound, geometry and property) in the teacher transfers well to the student.

Enhancer		Method			
KE	TE	GCN	GIN	GAT	GraphSAGE
×	×	66.88	66.68	66.53	66.99
×	✓	68.07 (1.19)	68.16 (1.48)	68.48 (1.95)	68.44 (1.45)
✓	×	68.26 (1.38)	68.60 (1.92)	68.59 (2.06)	68.58 (1.59)
✓	✓	69.81 (2.93)	70.16 (3.48)	69.76 (3.23)	69.61 (2.62)

Table 6: Ablation results on knowledge enhancer (KE) and task enhancer (TE). The average ROC-AUC (%) performance on 8 classification datasets is reported. The number in bracket indicates the absolute performance improvement compared to the baseline without KE and TE.

5 Conclusion

In this work, we propose a novel image-enhanced molecular graph representation learning framework (called IEM), which is the first attempt to use images to improve the performance of graphs. Equiped with a knowledgeable image-based teacher and 2 enhancers (knowledge enhancer and task enhancer), our IEM can significantly improve the performance of graph-based methods without any architectural modifications on a large number of drug discovery benchmarks. Therefore, IEM has the potential to leverage images to empower a wider range of graph representation learning fields, such as grid representation learning and skeleton representation learning. In particular, we experimently demonstrate that performance can be improved by cheap image rendering for microscopic entities where image data is difficult to obtain, which will encourage us to use IEM for more biological entities in life sciences, such as protein and ribonucleic acid.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant nos. 62122025, U22A2037, 62250028), Postgraduate Scientific Research Innovation Project of Hunan Province (grant no. CX20220380).

References

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021.

[Cohen *et al.*, 2009] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

[DeLano and others, 2002] Warren L DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, 40(1):82–92, 2002.

[Fernandez *et al.*, 2018] Michael Fernandez, Fuqiang Ban, Godwin Woo, Michael Hsing, Takeshi Yamazaki, Eric LeBlanc, Paul S Rennie, William J Welch, and Artem Cherkasov. Toxic colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images. *Journal of chemical information and modeling*, 58(8):1533–1543, 2018.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[Goh *et al.*, 2017] G. B. Goh, C. Siegel, A. Vishnu, N. O. Hodas, and N. Baker. Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed qsar/qspr models. 2017.

[Guo *et al.*, 2021] Zhichun Guo, Chuxu Zhang, Wenhao Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V Chawla. Few-shot graph learning for molecular property prediction. In *Proceedings of the Web Conference 2021*, pages 2559–2567, 2021.

[Gupta *et al.*, 2016] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Hou *et al.*, 2021] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5693–5702, October 2021.

[Hu *et al.*, 2017] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. Ogb-lsc: A large-scale challenge for machine learning on graphs. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2017.

[Hu *et al.*, 2020a] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *ICLR*, 2020.

[Hu *et al.*, 2020b] Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1857–1867, 2020.

[Jin *et al.*, 2023] Yufeng Jin, Guosheng Hu, Haonan Chen, Duoqian Miao, Liang Hu, and Cairong Zhao. Cross-modal distillation for speaker recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12977–12985, 2023.

[Landrum, 2013] Greg Landrum. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling, 2013.

[Li *et al.*, 2021a] Pengyong Li, Jun Wang, Yixuan Qiao, Hao Chen, Yihuan Yu, Xiaojun Yao, Peng Gao, Guotong Xie, and Sen Song. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Briefings in Bioinformatics*, 2021.

[Li *et al.*, 2021b] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 2021.

[Liang *et al.*, 2022] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[Likhosherstov *et al.*, 2021] Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv preprint arXiv:2111.12993*, 2021.

[Liu *et al.*, 2021] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

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

[Rong *et al.*, 2020] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In *NeurIPS*, 2020.

[Shen *et al.*, 2022] Boyuan Shen, Huiqiu Wang, Hao Xiong, Xiao Chen, Eric GT Bosch, Ivan Lazić, Weizhong Qian,

and Fei Wei. Atomic imaging of zeolite-confined single molecules by electron microscopy. *Nature*, 607(7920):703–707, 2022.

[Song *et al.*, 2022] Xingchen Song, Di Wu, Binbin Zhang, Zhiyong Wu, Wenpeng Li, Dongfang Li, Pengshen Zhang, Zhendong Peng, Fuping Pan, Changbao Zhu, et al. Fusionformer: Fusing operations in transformer for efficient streaming speech recognition. *arXiv preprint arXiv:2210.17079*, 2022.

[Stärk *et al.*, 2021] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lio. 3d infomax improves gnns for molecular property prediction. In *NeurIPS 2021 AI for Science Workshop*, 2021.

[Sun *et al.*, 2020] Fan-Yun Sun, Jordon Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *International Conference on Learning Representations*, 2020.

[Sun *et al.*, 2021] Jingxian Sun, Lichao Zhang, Yufei Zha, Abel Gonzalez-Garcia, Peng Zhang, Wei Huang, and Yanning Zhang. Unsupervised cross-modal distillation for thermal infrared tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2262–2270, 2021.

[Tumer and Ghosh, 1995] Kagan Tumer and Joydeep Ghosh. Theoretical foundations of linear and order statistics combiners for neural pattern classifiers. *IEEE Trans. Neural Networks*, 1995.

[Wang *et al.*, 2020] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 33:4835–4845, 2020.

[Wu *et al.*, 2018] 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.

[Wu *et al.*, 2020] 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*, 32(1):4–24, 2020.

[Xia *et al.*, 2023] Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*, 2023.

[Xiang *et al.*, 2023] Hongxin Xiang, Shuting Jin, Xiangrong Liu, Xiangxiang Zeng, and Li Zeng. Chemical structure-aware molecular image representation learning. *Briefings in Bioinformatics*, 24(6):bbad404, 2023.

[Xu *et al.*, 2018] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.

[Yan *et al.*, 2022] Xu Yan, Heshen Zhan, Chaoda Zheng, Jiantao Gao, Ruimao Zhang, Shuguang Cui, and Zhen Li. Let images give you more: Point cloud cross-modal training for shape analysis. *Advances in Neural Information Processing Systems*, 35:32398–32411, 2022.

[Yang *et al.*, 2021] Jianfei Yang, Jiangang Yang, Shizheng Wang, Shuxin Cao, Han Zou, and Lihua Xie. Advancing imbalanced domain adaptation: Cluster-level discrepancy minimization with a comprehensive benchmark. *IEEE Transactions on Cybernetics*, 2021.

[Zeng *et al.*, 2022] Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016, 2022.

[Zhang *et al.*, 2021] Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34, 2021.

[Zhang *et al.*, 2023] Xiang Zhang, Hongxin Xiang, Xixi Yang, Jingxin Dong, Xiangzheng Fu, Xiangxiang Zeng, Haowen Chen, and Keqin Li. Dual-view learning based on images and sequences for molecular property prediction. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[Zhong *et al.*, 2021] Shifa Zhong, Jiajie Hu, Xiong Yu, and Huichun Zhang. Molecular image-convolutional neural network (cnn) assisted qsar models for predicting contaminant reactivity toward oh radicals: Transfer learning, data augmentation and model interpretation. *Chemical Engineering Journal*, 408:127998, 2021.

[Zhou *et al.*, 2023] Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5116–5125, 2023.