Abstract

Graph property prediction tasks are important and numerous. While each task offers a small size of labeled examples, unlabeled graphs have been collected from various sources and at a large scale. A conventional approach is training a model with the unlabeled graphs on self-supervised tasks and then fine-tuning the model on the prediction tasks. However, the self-supervised task knowledge could not be aligned or sometimes conflicted with what the predictions needed. In this paper, we propose to extract the knowledge underlying the large set of unlabeled graphs as a specific set of useful data points to augment each property prediction model. We use a diffusion model to fully utilize the unlabeled graphs and design two new objectives to guide the model’s denoising process with each task’s labeled data to generate task-specific graph examples and their labels. Experiments demonstrate that our data-centric approach performs significantly better than fifteen existing various methods on fifteen tasks. The performance improvement brought by unlabeled data is visible as the generated labeled examples unlike the self-supervised learning.

1 Introduction

Graph data such as molecules and polymers are found to have attractive properties in drug and material discovery (Böhm et al., 2004; Huang et al., 2021), but annotating them requires specialized knowledge, as well as lengthy and costly experiments in wet labs (Cormack and Elorza, 2004). So, it is important for graph property predictors to learn useful knowledge from unlabeled graphs.

Self-supervised learning (Hu et al., 2019; Rong et al., 2020; You et al., 2021; Kim et al., 2022) utilizes unlabeled graphs to learn through predictive tasks or contrastive tasks to represent and transfer the knowledge as model parameters. Despite the empirical success in language and vision (Brown et al., 2020; He et al., 2022), their performance on graph data applications remains unsatisfactory because of the significant gap between the graph self-supervised task and the graph label prediction task. Models trained on node attribute prediction (Hu et al., 2019) as a simple predictive self-supervised task extract too limited knowledge from the graph structure, which has been observed after too fast convergence (Sun et al., 2022). More complex tasks like graph context prediction (Hu et al., 2019; Zhang et al., 2021) may transfer knowledge that conflicts with downstream tasks. Aromatic rings, for instance, are a prevalent structure in molecules (Maziarka et al., 2020) and are considered valuable in context prediction tasks (Zhang et al., 2021). However, graph properties such as oxygen permeability can be more related to non-aromatic rings in some cases (Liu et al., 2022a), which is overlooked if not using tailored predictive tasks specifically for downstream tasks. As predictive tasks strive for
Existing approach: Knowledge from self-supervised tasks could not be aligned or even conflict with what predictions need. Parameter initialization could hardy interpret how unlabeled graphs were or would be able to improve the models, leading to high prediction errors.

Data-centric approach: Target knowledge in labeled graphs guides the denoising process in diffusion model to generate new labeled examples close to the target graph space instead of the unlabeled graph space. The augmented knowledge for the prediction model is visible as graphs.

Our data-centric idea avoids the use of self-supervised tasks that are not appropriate. We use a diffusion probabilistic model (known as diffusion model) to capture the data distribution of unlabeled graphs, leveraging its capability of distribution coverage, stationarity, and scalability (Dhariwal and Nichol 2021). At the stage of performing a particular property prediction task, the reverse process, guided by novel task-related optimization objectives, generates new task-specific labeled examples. Minimal sufficient knowledge from the unlabeled data is transferred into these examples, instead of uninterpretable model parameters, and then to enhance the training of prediction models.

To implement our idea, we propose a Data-Centric Transfer framework (DCT) based on a diffusion model for graph data, as shown in Figure 1b. It aims to transfer minimal sufficient knowledge from unlabeled graphs to property predictors by data augmentation. The diffusion model gradually adds Gaussian noise to a graph from which a score function (i.e., the gradient of the log probability density) is then learned to estimate the noise step by step to reverse the process. DCT trains the diffusion model on the unlabeled graphs to get ready to augment any labeled dataset. Given a labeled graph from a particular task (i.e., type of property), the diffusion model adds noise to perturb it by a few steps and then generates a new graph through the score function. The new graph could be close to the distribution of the unlabeled graphs for diversity, however, it would lose the relatedness to the target task. So, we add two task-related objectives into the score function to guide the reverse process. When a predictor model $f$ has been trained on the task, given an original labeled graph $G$, the first objective is to optimize the new graph $G'$ to sufficiently preserve the label of $G$ with $f$. The second objective is to optimize $G'$ to be very different from (i.e., minimally similar to) $G$. These two objectives ensure that $G'$ carries minimal sufficient knowledge from the unlabeled graphs to be an augmentation of $G$. DCT iteratively generates new examples to augment the labeled dataset and progressively trains the prediction model with it.

We test DCT on fifteen graph property prediction datasets from three fields: chemistry (molecules), material science (polymers), and biology (protein-protein interaction graphs). DCT achieves the best performance over all these tasks. We find that the state-of-the-art self-supervised methods often struggle to transfer knowledge to regression tasks, etc. DCT reduces the mean absolute error relatively by 13.4% and 10.2% compared to the best baseline on the molecule and polymer graph regression tasks, respectively.

Figure 1: Comparing the diagrams of the existing approach and the proposed approach to learning from unlabeled graphs for a variety of graph property prediction tasks.
2 Problem Definition

Given $K$ property prediction tasks, there are $N^{[k]}$ labeled graph examples for the $k$-th task. They are \{(G_i, y_i) | G_i \in G^{[k]}, y_i \in \mathcal{Y}^{[k]}\}_{i=1}^{N^{[k]}},$ where $G^{[k]}$ is the graph space and $\mathcal{Y}^{[k]}$ is the label space of the task. The prediction model with parameters $\theta$ is defined as $f^{[k]}_{\theta} : G^{[k]} \rightarrow \mathcal{Y}^{[k]}$. $f^{[k]}_{\theta}$ consists of a GNN and a multi-layer perceptron (MLP). Without the loss of generality, we consider Graph Isomorphism Networks (GIN) (Xu et al., 2019) to encode graph structures. Given a graph $G = (\mathcal{V}, \mathcal{E}) \in G^{[k]}$ in the task $k$, GIN updates the representation vector of node $v \in \mathcal{V}$ at $l$-layer:

$$h^l_v = \text{MLP}^l\left((1 + \epsilon) \cdot h^{l-1}_v + \sum_{u \in \mathcal{N}(v)} h^{l-1}_u\right),$$

where $\epsilon$ is a learnable scalar and $u \in \mathcal{N}(v)$ is one of node $v$’s neighbor nodes. After stacking $L$ layers, the READOUT function (e.g., summation) gets the graph representation across all the nodes. The predicted label is:

$$\hat{y} = \text{MLP} \left(\text{READOUT} \left(\{h^L_v | v \in G\}\right)\right).$$

$f^{[k]}_{\theta}$ is hard to be well-trained because it is hard to collect graph labels at a large scale ($N^{[k]}$ is small). Fortunately, regardless of the tasks, a large number of unlabeled graphs are usually available from the same or similar domains. Self-supervised learning methods (Hu et al., 2019) rely on hand-crafted tasks to extract knowledge from the unlabeled examples $\{G_j \in G^{[T]}, j = 1, \ldots, N\}$ as pre-trained model parameters $\theta$. The uninterpretable parameters are transferred to warm up the prediction models $\{f^{[k]}_{\theta}\}_{k=1}^K$ on the $K$ downstream graph property prediction tasks. However, the gap and even conflict between the self-supervised tasks and the property prediction tasks lead to suboptimal performance of the prediction models. In the next section, we present the DCT framework that transfers knowledge from the unlabeled graphs with a data-centric approach.

3 The Data-Centric Transfer Framework

3.1 Overview of Proposed Framework

The goal of data-centric approaches is to augment training datasets by generating useful labeled data examples. Under that, the goal of our data-centric transfer (DCT) framework is to transfer the knowledge from unlabeled data into the data augmentation. Specifically, for each graph-label pair $(G^{[k]} \in G^{[T]}, y^{[k]} \in \mathcal{Y}^{[k]})$ in the task $k$, the framework is expected to output a new example $G'^{[k]}$ with the label $y'^{[k]}$ such that (1) $y'^{[k]} = y^{[k]}$ and (2) $G'^{[k]}$ and $G^{[k]}$ are from the same graph space $G^{[k]}$. However, if the graph structures of $G'^{[k]}$ and $G^{[k]}$ were too similar, the augmentation would duplicate the original data examples, become useless, and even cause over-fitting. So, the optimal graph data augmentation should enrich the training data with good diversity as well as preserve the labels of the original graphs. To achieve this, DCT utilizes a diffusion probabilistic model to first learn the data distribution from unlabeled graphs (Section 3.3). Then DCT adapts the reverse process in the diffusion model to generate task-specific labeled graphs for data augmentation (Section 3.3). Thus, the augmented graphs will be derived from the distribution of a huge collection of unlabeled data for diversity. To preserve the labels, DCT controls the reverse process with two task-related optimization objectives to transfer minimal sufficient knowledge from the unlabeled data. The first objective minimizes an upper bound of mutual information between the augmented and the original graphs in the graph space. The second objective maximizes the probability of the predicted label of augmented graphs being the same as the label of original graphs. The first is for minimal knowledge transfer, and the second is for sufficient knowledge transfer. DCT integrates the two objectives into the reverse process of the diffusion model to guide the generation of new labeled graphs. DCT iteratively trains the graph property predictor (used in the second objective) and creates the augmented training data. To simplify notations, we remove the task superscript $[k]$ in the following sections.

3.2 Learning Data Distribution from Unlabeled Graphs

The diffusion process for graphs in Figure 2 applies to both graph structure and node features. The diffusion model slowly corrupts unlabeled graphs to a standard normal distribution with noise. For
In this section, we focus on task-specific data augmentation using objectives \( I_1 \) and \( I_2 \) in the reverse process. The model was trained on unlabeled graphs to learn the general data distribution. Then it generates \( (G', y' = y) \) based on \( (G, y) \) in the reverse process. It perturbs \( G \) with \( D \) steps and optimizes \( G' \) to be minimally similar to \( G \) (Objective \( I_1 \)) and sufficiently preserve the label of \( G \) (Objective \( I_2 \)).

**3.3 Generating Task-specific Labeled Graphs**

Self-training approaches would propose to either (1) select unlabeled graphs by a graph property predictor or (2) generate graphs directly from the standard diffusion model, and then use the predictor to assign them labels so that the training data could be enriched. However, we have observed that neither of them can guarantee positive impact on the prediction performance. In fact, as shown in Figure 3, they make very little or even negative impact. That is because the selected or directly-generated graphs are too different from the labeled graph space of the target tasks. Task details of ten datasets are in appendix C.
Given a labeled graph \((G, y)\) from the original dataset of a specific task, the new labeled graph \((G', y')\) is expected to provide useful knowledge to augment the training set. We name it the augmented graph throughout this section. The augmented graph is desired to have the following two properties, as in Section 3.1 Task relatedness: As an effective training data point, \(G' \in \mathcal{G}\) and \(y' \in \mathcal{Y}\) are from the graph/label spaces of the specific task where \((G, y)\) come from and thus transfer sufficient task knowledge into the training set; Diversity: If \(G'\) was too similar to \(G\), the new data point would cause severe over-fitting on the property prediction model. The augmentation aims to learn from unlabeled graph to create diverse data points, which should contain minimal task knowledge about \(G\).

The selected unlabeled graphs used in self-training have little task relatedness because the unlabeled data distribution might be too far from the one of the specific task. Existing graph data augmentation methods could not create diverse graph examples because they manipulated labeled graphs and did not learn from the unlabeled graphs. Our novel data-centric approach DCT works towards both desired properties by transferring minimally sufficient knowledge from the unlabeled graphs: Sufficiency is achieved by maximizing the possibility for label preservation (i.e., \(y' = y\)). It ensures that the knowledge from unlabeled graphs is task-related; Minimality refers to the minimization of graph similarity between \(G'\) and \(G\) to ensure that the augmentation introduces diversity. Both optimizations can be formulated using mutual information \(\mathcal{I}(\cdot; \cdot)\) to generate task-specific labeled data \((G', y')\).

**Definition 3.1** (Sufficiency for Data Augmentation). The augmented graph \(G'\) sufficiently preserves the label of the original graph \(G\) if and only if \(\mathcal{I}(G'; G) \leq \mathcal{I}(G; G), \forall G\) represents any augmented graph that sufficiently preserves the original graph’s label.

Self-supervised tasks applied a similar philosophy in pre-training \cite{soatto2016unsupervised}, however, they did not use labeled data from any specific tasks. So the optimizations were unable to extract useful knowledge and transfer it to the downstream \cite{tian2020does}. In our DCT that performs task-specific data augmentation, the augmented graphs can be optimized toward the objectives using any labeled graph \(G\) and its label \(y\):

\[
\min_{\mathcal{I}_1} \max_{\mathcal{I}_2} \mathbb{E}_G [\mathcal{I}_1 (G'; G) + \mathcal{I}_2 (G'; y)].
\]

For the first objective, we use the leave-one-out variant of InfoNCE \cite{pmlr-v80-poole18a, oord2018representation} as the upper bound estimation. For the \(i\)-th labeled graph \((G_i, y_i)\),

\[
\mathcal{I}_1 \leq \mathcal{I}_{\text{bound}} (G_i'; G_i) = \log \frac{p(G_i'|G_i)}{\sum_{j=1,j\neq i}^M p(G_i'|G_j)},
\]

where \(G_i'\) is the augmented graph. When \(G_i'\) is optimized, \(G_i\) makes a positive pair; \(\{G_j\} (j \neq i)\) are \(M - 1\) negative samples of labels that do not equal \(y_i\). \((M\) is a hyperparameter). We use cosine similarity and a softmax function to calculate \(p(G_i'|G_j) = \frac{\exp(sim(G_i', G_j))}{\sum_{j=1}^M \exp(sim(G_i', G_j))}\). In practice, we extract statistical features of graphs to calculate their similarity. Details are in appendix B.2.

For the second objective, we denote the predicted label of the augmented graph \(G'\) by \(f_\theta(G')\). We maximize the log likelihood \(\log p(y | f_\theta(G'))\) to maximize \(\mathcal{I}_2(G'; y)\). Specifically, after the predictor \(f_\theta\) is trained for several epochs on the labeled data, we freeze its parameters and use it to optimize the augmented graphs so they are task-related:

\[
\mathcal{L}(G') = \mathcal{I}_{\text{bound}} (G'; G) - \log p(y | f_\theta(G')).
\]

**Framework details**: As shown in Figure 2, after the diffusion model learns the data distribution from unlabeled graphs, given a labeled graph \(G\) from a specific task, DCT perturbs it for \(D (D \ll T)\) steps. The perturbed noisy graph, denoted by \(\tilde{G}(t)\), stays inside the task-specific graph and label space, rather than the noise distribution (at step \(T\)). To reverse the noise in it and generate a task-specific augmented example \(G'\), DCT integrates the loss function in Eq. (7) into the score function \(s(\cdot, t)\) for minimal sufficient knowledge transfer:

\[
d\tilde{G}(t) = f(\tilde{G}(t), t) - g(t)^2 \left( s(\tilde{G}(t), t) - \alpha \nabla_{\tilde{G}(t)} \mathcal{L}(\tilde{G}(t)) \right) dt + g(t)d\mathbf{w},
\]

where \(\alpha\) is a scalar for score alignment between \(s\) and \(\nabla \mathcal{L}\) to avoid the dominance of any of them:

\[
\alpha = \frac{\|s(\tilde{G}(t), t)\|_2}{\|\nabla_{\tilde{G}(t)} \mathcal{L}(\tilde{G}(t))\|_2}.
\]

Because \(\tilde{G}(t)\) is an intermediate state in the reverse process, the noise in it
Table 1: Statistics of datasets for graph property prediction in different domains.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Dataset</th>
<th># Graphs</th>
<th>Prediction Task</th>
<th># Task</th>
<th>Avg./Max # Nodes</th>
<th>Avg./Max # Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecules</td>
<td>ogbg-HIV</td>
<td>41,127</td>
<td>Classification</td>
<td>1</td>
<td>25.5 / 222</td>
<td>54.9 / 502</td>
</tr>
<tr>
<td></td>
<td>ogbg-ToxCast</td>
<td>8,576</td>
<td>Classification</td>
<td>617</td>
<td>18.8 / 124</td>
<td>38.5 / 268</td>
</tr>
<tr>
<td></td>
<td>ogbg-Tox21</td>
<td>7,831</td>
<td>Classification</td>
<td>12</td>
<td>18.6 / 132</td>
<td>38.6 / 290</td>
</tr>
<tr>
<td></td>
<td>ogbg-BBBP</td>
<td>2,039</td>
<td>Classification</td>
<td>1</td>
<td>24.1 / 132</td>
<td>51.9 / 290</td>
</tr>
<tr>
<td></td>
<td>ogbg-BACE</td>
<td>1,513</td>
<td>Classification</td>
<td>1</td>
<td>34.1 / 97</td>
<td>73.7 / 202</td>
</tr>
<tr>
<td></td>
<td>ogbg-ChinTox</td>
<td>1,477</td>
<td>Classification</td>
<td>2</td>
<td>26.2 / 136</td>
<td>55.8 / 286</td>
</tr>
<tr>
<td></td>
<td>ogbg-SIDER</td>
<td>1,427</td>
<td>Classification</td>
<td>12</td>
<td>33.6 / 492</td>
<td>70.7 / 1010</td>
</tr>
<tr>
<td></td>
<td>ogbg-Lipo</td>
<td>4200</td>
<td>Regression</td>
<td>1</td>
<td>27 / 115</td>
<td>59 / 236</td>
</tr>
<tr>
<td></td>
<td>ogbg-ESOL</td>
<td>1128</td>
<td>Regression</td>
<td>1</td>
<td>13.3 / 55</td>
<td>27.4 / 124</td>
</tr>
<tr>
<td></td>
<td>ogbg-FreeSolv</td>
<td>642</td>
<td>Regression</td>
<td>1</td>
<td>8.7 / 24</td>
<td>16.8 / 50</td>
</tr>
<tr>
<td>Polymers</td>
<td>GlassTemp</td>
<td>7,174</td>
<td>Regression</td>
<td>1</td>
<td>36.7 / 166</td>
<td>79.3 / 362</td>
</tr>
<tr>
<td></td>
<td>MeltingTemp</td>
<td>3,651</td>
<td>Regression</td>
<td>1</td>
<td>26.9 / 102</td>
<td>55.4 / 212</td>
</tr>
<tr>
<td></td>
<td>ThermCond</td>
<td>759</td>
<td>Regression</td>
<td>1</td>
<td>21.3 / 71</td>
<td>42.3 / 162</td>
</tr>
<tr>
<td></td>
<td>O2Perm</td>
<td>595</td>
<td>Regression</td>
<td>1</td>
<td>37.3 / 103</td>
<td>82.1 / 234</td>
</tr>
<tr>
<td>Proteins</td>
<td>PPI</td>
<td>88000</td>
<td>Classification</td>
<td>40</td>
<td>49.4 / 111</td>
<td>890.8 / 11556</td>
</tr>
</tbody>
</table>

may fail the optimizations. So, we design a new sampling method named double-loop sampling for accurate loss calculation. It has an inner-loop sampling using Eq. (4) to sample $\hat{G}(t)$, as the denoised version of $\tilde{G}(t)$ at the reverse time $t$. Then $\nabla_{\hat{G}} \mathcal{L}(\hat{G}(t))$ is calculated as an alternative for $\nabla_{\tilde{G}} \mathcal{L}(\tilde{G}(t))$. Finally, an outer-loop sampling takes one step to guide denoising using Eq. (8).

DCT iteratively creates the augmented graphs $(G', y')$, updates the training dataset $\{ (G_i, y_i) \}$, and trains the graph property predictor $f_\theta$. In each iteration, for task $k$, $n \ll N[k]$ labelled graphs of the lowest property prediction loss are selected to create the augmented graphs. The predictor is better fitted to these graphs for more accurate sufficiency estimation of the augmentation.

4 Experiments

In this section, we present and analyze experimental results to demonstrate the outstanding performance of DCT, the usefulness of new optimization objectives, the effect of hyperparameters and iterative process, and the interpretability of “visible” knowledge transfer from unlabeled graphs.

4.1 Experimental Setup

Tasks and metrics: Experiments are conducted on 15 graph property prediction tasks in chemistry, material science, and biology, including seven molecule classification, three molecule regression tasks from open graph benchmarks (Hu et al., 2020), four polymer regression tasks, and protein function prediction (PPI) (Hu et al., 2019). Dataset statistics is presented in Table 1. We use the area under the ROC curve (AUC) to evaluate classifiers and mean absolute error (MAE) for regressors.

Baselines and implementation: Besides GIN, there are three lines of baseline methods: (1) self-supervised learning methods including EdgePred, AttrMask, ContextPred in (Hu et al., 2019), InfoMax (Velickovic et al., 2019), JOAO (You et al., 2021), GraphLOG (Xu et al., 2021), MGSSL (Zhang et al., 2021) and D-SLA (Kim et al., 2022), (2) semi-supervised learning methods including self-training with selected unlabeled graphs (ST-REAL) and generated graphs (ST-GEN) and INFOGRAPH (Sun et al., 2020), and (3) graph data augmentation (GDA) methods including FLAG (Kong et al., 2022), GREAX (Liu et al., 2022a), and G-MixUp (Han et al., 2022). For self-supervised pre-training, we follow their own settings and directly use their pre-trained models if available. For semi-supervised learning methods and DCT, we use 113K QM9 (Ramakrishnan et al., 2014) and 306K PPI graphs (Hu et al., 2019) as unlabeled data sources for the tasks on molecules/polymers and proteins, respectively. For DCT, we tune three major hyper-parameters: the number of perturbation steps $D \in [1, 10]$, the number of negative samples $M \in [1, 10]$, and top-$n$ % labeled graphs of lowest property prediction loss selected for data augmentation.

4.2 Outstanding Property Prediction Performance

We report the model performance using mean and standard deviation over 10 runs Table 2. DCT is the best on all 15 tasks compared to the state-of-the-art baselines. Our observations are:
Table 2: Mean(±σ) on tasks from different fields. The best mean is **bold**. The best baseline is underlined. Results are highlighted if unlabeled graphs bring significant negative impacts compared to GIN. The MAE for ThermCond is scaled × 100. G-MIXUP was proposed for classification. MGSSL was proposed for molecules.

<table>
<thead>
<tr>
<th>Molecule Classification: AUC (%)</th>
<th>ogbg-HIV</th>
<th>ogbg-ToxCast</th>
<th>ogbg-Tox21</th>
<th>ogbg-BBPP</th>
<th>ogbg-BACE</th>
<th>ogbg-ClinTox</th>
<th>ogbg-SIDER</th>
</tr>
</thead>
<tbody>
<tr>
<td># Training Graphs</td>
<td>32,901</td>
<td>6,860</td>
<td>6,264</td>
<td>1,631</td>
<td>1,210</td>
<td>1,181</td>
<td>1,141</td>
</tr>
<tr>
<td>GIN</td>
<td>77.4(1.2)</td>
<td>66.9(0.2)</td>
<td>76.0(0.6)</td>
<td>67.5(2.7)</td>
<td>77.5(2.8)</td>
<td>88.8(3.8)</td>
<td>58.1(0.9)</td>
</tr>
<tr>
<td>EDGEPred</td>
<td>78.1(1.3)</td>
<td>63.9(0.4)</td>
<td>75.5(0.4)</td>
<td>69.9(0.5)</td>
<td>79.5(1.0)</td>
<td>62.9(2.3)</td>
<td>59.7(0.8)</td>
</tr>
<tr>
<td>ATTRMark</td>
<td>77.1(1.7)</td>
<td>64.2(0.5)</td>
<td>76.6(0.4)</td>
<td>63.9(1.2)</td>
<td>79.3(0.7)</td>
<td>70.4(1.1)</td>
<td>60.7(0.4)</td>
</tr>
<tr>
<td>CONTEXTPred</td>
<td>78.4(0.1)</td>
<td>63.7(0.3)</td>
<td>75.0(0.3)</td>
<td>68.1(1.6)</td>
<td>75.7(1.0)</td>
<td>63.2(1.5)</td>
<td>68.0(0.8)</td>
</tr>
<tr>
<td>InfoMAX</td>
<td>75.4(1.8)</td>
<td>61.7(0.1)</td>
<td>75.5(0.4)</td>
<td>69.2(0.5)</td>
<td>76.8(0.2)</td>
<td>73.0(0.2)</td>
<td>58.0(0.5)</td>
</tr>
<tr>
<td>JOAO</td>
<td>76.2(0.2)</td>
<td>64.8(0.3)</td>
<td>74.8(0.5)</td>
<td>69.3(2.5)</td>
<td>75.9(3.9)</td>
<td>69.4(5.5)</td>
<td>68.0(0.4)</td>
</tr>
<tr>
<td>GraphLogOM</td>
<td>74.8(1.1)</td>
<td>63.2(0.8)</td>
<td>75.4(0.8)</td>
<td>67.5(2.3)</td>
<td>80.4(3.6)</td>
<td>69.0(8.0)</td>
<td>57.0(5.9)</td>
</tr>
<tr>
<td>MGSSL</td>
<td>71.1(1.1)</td>
<td>65.7(0.4)</td>
<td>77.2(0.3)</td>
<td>66.9(0.9)</td>
<td>81.3(2.4)</td>
<td>69.8(5.0)</td>
<td>63.8(1.0)</td>
</tr>
<tr>
<td>D-SLA</td>
<td>76.9(0.9)</td>
<td>60.8(1.2)</td>
<td>76.1(1.1)</td>
<td>62.6(1.0)</td>
<td>80.3(0.6)</td>
<td>78.2(2.1)</td>
<td>55.1(1.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Self-Supervised</th>
<th>INFOGRAPH</th>
<th>ST-REAL</th>
<th>ST-GEN</th>
<th>DCT (Ours)</th>
<th>GDA</th>
<th>GREA</th>
<th>G-MIXUP</th>
</tr>
</thead>
<tbody>
<tr>
<td># Training Graphs</td>
<td>60,361</td>
<td>60,812</td>
<td>61,141</td>
<td>76,109</td>
<td>60,812</td>
<td>68,101</td>
<td>68,101</td>
</tr>
<tr>
<td>MAE</td>
<td>73.3(0.7)</td>
<td>61.5(1.3)</td>
<td>76.0(0.9)</td>
<td>76.1(1.0)</td>
<td>67.9(0.7)</td>
<td>79.1(1.2)</td>
<td>82.2(2.4)</td>
</tr>
<tr>
<td>AMPRE</td>
<td>78.3(0.6)</td>
<td>65.4(1.0)</td>
<td>76.2(0.5)</td>
<td>66.7(1.9)</td>
<td>77.4(1.8)</td>
<td>82.2(2.4)</td>
<td>60.8(1.2)</td>
</tr>
<tr>
<td>ST-R</td>
<td>77.9(1.6)</td>
<td>65.1(0.0)</td>
<td>75.8(0.9)</td>
<td>66.3(1.5)</td>
<td>78.4(0.0)</td>
<td>87.3(1.3)</td>
<td>59.3(1.3)</td>
</tr>
</tbody>
</table>

(1) **GIN** is the most competitive baseline and outperforms self-supervised learning methods. On 7 of 15 tasks, GIN outperforms all the 7 self-supervised learning methods. Because self-supervised pre-training imposes constraints on the model architecture, it undermines the true power of GNNs and under-performs the GNNs that are properly used.

(2) **Self-training and GDA methods perform better than GIN but cannot effectively learn from unlabeled data.** Self-training (ST-REAL and ST-GEN) is often the best baseline in regression tasks. GDA (GREA and G-MIXUP) methods outperform self-training in most classification tasks except ogbg-SIDER, because they are often designed to exploit categorical labeled data and remain under-performance for regression. Although self-training benefits from selecting unlabeled examples in some graph regression tasks, they are negatively affected by the unlabeled graphs in the classification tasks such as ogbg-ToxCast and ogbg-ClinTox. As indicated in Figure 3, it is inappropriate to pseudo-label unlabeled graphs in self-training due to the huge gap between the unlabeled data and target task.

(3) **DCT transfers useful knowledge from unlabeled data by data augmentation.** DCT outperforms the best baseline relatively by +3.9%, +13.4%, and +10.2% when there are only 1,210, 513, and 4,303 training graphs on ogbg-BACE, ogbg-FreeSolv, and GlassTemp, respectively. Compared to the self-supervised baselines, the improvement from DCT is more significant, so the knowledge transfer is more effective. For example, on ogbg-FreeSolv and O2-Perm, DCT performs better than the self-supervised baselines relatively by +45.8% and +8.0%, respectively. On regression tasks that involve knowledge transfer across domains (e.g., from molecules to polymers), DCT reduces MAE relatively by 1.9% ~ 10.2% compared to the best baseline. All these results demonstrate the outstanding performance of task-specific data augmentation in DCT.
Table 3: Comprehensive ablation studies for DCT on tasks ogbg-BACE, ogbg-SIDER, ogbg-FreeSolv, and O₂Perm. Objectives include minimizing \( \mathcal{I}_1(G', G) \) and/or maximizing \( \mathcal{I}_2(G', y) \).

<table>
<thead>
<tr>
<th>Objectives</th>
<th>Classification</th>
<th>Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{I}_1(G', G) )</td>
<td>BACE</td>
<td>SIDER</td>
</tr>
<tr>
<td>Top Baseline Method</td>
<td>( \times )</td>
<td>( \times )</td>
</tr>
<tr>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
<td>85.2(1.13)</td>
</tr>
<tr>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
<td>84.7(1.84)</td>
</tr>
<tr>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
<td>85.6(0.6)</td>
</tr>
<tr>
<td>Unlabeled Data Sources</td>
<td>( \times )</td>
<td>( \times )</td>
</tr>
<tr>
<td>QM9</td>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>ZINC</td>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
<td>84.9(0.4)</td>
</tr>
</tbody>
</table>

Figure 4: Effect of hyper-parameters, including the number of perturbation steps \( D \in \{1, 10\} \), the number of negative graphs \( M \in \{1, 10\} \), and top-\( n \% \) labeled graphs whose labels are predicted the most accurately and that are selected for data augmentation, where \( n \in \{10, 100\} \).

4.3 Ablation Studies and Performance Analysis

**Comprehensive ablation studies:** In Table 3, we investigate how the task-related objectives in Eq. 5 impact the performance of DCT. First, DCT outperforms the top baseline even if the two task-related optimization objectives are disabled. This is because DCT generates new training examples based on original labeled graphs: the data augmentation has already improved the diversity of the training dataset a little bit. Second, adding the objective \( \mathcal{I}_2 \) further improves the performance by encouraging the generation of diverse examples, because it minimizes the similarity between the original graph and augmented graph in the graph space. Third, we receive the best performance of DCT when it combines \( \mathcal{I}_1 \) and \( \mathcal{I}_2 \) objectives to generate task-related and diverse augmented graphs. When we change the unlabeled data source from QM9 to the ZINC dataset from \( \text{Jo et al., 2022} \), similar observations confirm the necessity of the task-related objectives.

**Effect of hyper-parameters:** The impacts of three hyper-parameters of DCT are studied: the number of perturbation steps \( D \), the number of negative samples \( M \) in Eq. (6), and the number of augmented graphs in each iteration (i.e., top-\( n \% \) selected graph for augmentation). Results from Figure 4 show that DCT is robust to a wide range of \( D \) and \( M \) valued from 0 to 10. They suggest that \( D \) and \( M \) can be set as 5 in most cases. As for the number of the augmented graphs in each iteration, results show that noisy graphs are often created when \( n \) is higher than 30\%, because the predictor cannot effectively guide the data augmentation for those labeled graphs whose labels are hard to predict. So, 10\% is suggested as the default of top-\( n \% \).

**Iterative process:** Figure 5 investigates the relationship between the quality of augmented graphs and the accuracy of property prediction models. We save a predictor checkpoint every 20 epochs.
to guide the generation of the augmented examples. We evaluate the quality of augmented graphs by using them to train GIN and report AUC/MAE. The data augmentation gradually decreases the training loss of property prediction. On the other hand, the increased GIN performance indicates that the quality of augmented examples is also improved over epochs. The data augmentation and predictor training mutual enhance each other.

4.4 Interpretability of Visible Knowledge Transfer

Knowledge transfer by data augmentation gives visible examples, allowing us to study what is learned. We visualize a few augmented graphs in DCT using ogbg-BACE and O$_2$Perm. We adapt top-k pooling [Knyazev et al., 2019] to select the subgraphs that GIN used for prediction. The selected subgraphs are highlighted in green in Figure 6. The three examples show that the augmented graphs can identify and preserve the core structures that GIN uses to predict property values. These augmented graphs are chemically valid, showing that concepts such as some chemical rules from the unlabeled graphs are successfully transferred to downstream tasks. More results are in appendix D.2.

Regarding task-specific knowledge, it is known that the fluorine atom and the methyl group are usually negatively and positively correlated to the permeability, respectively [Park et al., 2003; Corrado and Guo, 2020]. The augmented examples show that DCT captures this domain knowledge with the task-related objectives. In example (b), DCT replaces most of the fluorine atoms with the methyl groups. It encourages GIN to learn the positive relationship between the methyl group and the permeability so that GIN predicts a high label value. In example (c), DCT replaces the methyl groups with fluorine atoms. It encourages GIN to learn the negative relationship between the fluorine atom and the permeability so that GIN predicts a low label value.

5 Related Work

5.1 Graph Property Prediction

Graph neural networks (GNNs) [Kipf and Welling, 2017; Xu et al., 2019] are commonly used for graph property prediction in chemistry and polymer informatics tasks [Otsuka et al., 2011; Hu et al., 2020; Zhou et al., 2022]. However, it is hard to annotate enough labels in these domains. Recent work used self-supervised tasks such as node attribute prediction and graph structure prediction [Hu et al., 2019; You et al., 2021; Kim et al., 2022] to pre-train architecture-fixed GNNs. Sun et al. [2022] observed that the existing methods might fail to transfer knowledge from unlabeled graph data. Flexible GNN architectures for downstream tasks would be desirable.
Graph data augmentation (GDA) methods do not restrict GNN architecture choices to improve prediction accuracy (Trivedi et al., 2022; Zhao et al., 2022, 2023; Ding et al., 2022). They learn to create new examples that preserve the properties of original graphs (Liu et al., 2022b; Kong et al., 2022; Han et al., 2022; Luo et al., 2022). However, they purely manipulate labeled examples and thus cannot utilize the knowledge in unlabeled graphs. Our DCT combines the knowledge from the unlabeled dataset and the labeled task dataset. It creates label-preserved graph examples with the knowledge transferred from the unlabeled data. It allows the GNN models to have flexible architectures.

5.2 Learning from Unlabeled Data

Pre-training on self-supervised tasks such as masked image modeling and autoregressive text generation is effective for large language and vision models (Brown et al., 2020; He et al., 2022). However, the hand-crafted self-supervised tasks could hardly help models learn useful knowledge from unlabeled graphs due to the gap between these label-agnostic tasks and the downstream prediction tasks towards drug discovery and material discovery (Sun et al., 2021; Kim et al., 2022; Inae et al., 2023). A universal self-supervised task to learn from the unlabeled graphs remains under-explored (Sun et al., 2022; Trivedi et al., 2022).

Semi-supervised learning assumes that unlabeled and labeled data are from the same source (Liu et al., 2023). The learning objective in the latent space is usually mutual information maximization that encourages similarity between the representations of unlabeled and labeled graphs (Sun et al., 2020). However, the distributions of the unlabeled and labeled data could be very different due to the different types of sources (Hu et al., 2019), leading to negative impacts on the property prediction on the labeled graphs. Self-training, as a specific type of semi-supervised learning method, selects the unlabeled graphs of confidently predictable labels and assigns pseudo-labels for them (Lee et al., 2013; Iscen et al., 2019). Many studies have explored improving uncertainty estimation (Gal and Ghahramani, 2016; Tagasovska and Lopez-Paz, 2019; Amini et al., 2020) to help the model filter out noise for reliable pseudo-labels. Recently, pseudo-labels have been applied in imbalanced learning (Liu et al., 2023) and representation learning (Ghiasi et al., 2021). However, self-training is restricted to confidently predictable labels and may ignore the huge number of any other unlabeled graphs (Huang et al., 2022). Therefore, it cannot fully utilize the knowledge in the unlabeled graphs.

In contrast, our DCT employs a diffusion model to extract knowledge (as the diffusion and reverse processes) from all the unlabeled graphs. DCT represents the knowledge as task-specific labeled examples to augment the target dataset, instead of uninterpretable pre-trained model parameters. We note that self- or semi-supervised learning does not conflict with DCT, and we leave their combinations for future work.

5.3 Diffusion Models on Graphs

Recent works have improved the diffusion models on graphs (Niu et al., 2020; Jo et al., 2022; Vignac et al., 2022; Kong et al., 2023; Chen et al., 2023). EDP-GNN (Niu et al., 2020) employed score matching for permutation-invariant graph data distribution. GDSS (Jo et al., 2022) extended the continuous-time framework [6] to model node-edge joint distribution. DiGress (Vignac et al., 2022) used the transition matrix to preserve the discrete natures of the graph structure. GraphARM (Kong et al., 2023) introduced a node-absorbing autoregressive diffusion process. EDGE (Chen et al., 2023) focused on efficiently generating larger graphs. Instead of improving the generation performance of the diffusion model, our model builds on score-based diffusion models (Jo et al., 2022; Song et al., 2021) for predictive tasks, i.e., graph classification and graph regression.

6 Conclusion

In this work, we made the first attempt to transfer minimal sufficient knowledge from unlabeled graphs by data augmentation. We proposed a data-centric framework to use the diffusion model trained on the unlabeled graphs and use two task-related objectives to generate task-specific augmented graphs. Experiments demonstrated the performance of the proposed framework through visible augmented examples. It is better than self-supervised learning, self-training, and graph data augmentation methods on as many as 15 tasks.
Acknowledgments and Disclosure of Funding

This work was supported by NSF IIS-2142827, IIS-2146761, IIS-2234058, CBET-2102592, and ONR N00014-22-1-2507.

References


Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16000–16009, 2022.


A Additional Related Work on Data-Centric Approach

Data Augmentation  Data augmentation creates new examples with preserved labels but uses no unlabeled data (Shorten and Khoshgoftaar 2019; Kashefi and Hwa 2020; Balestriero et al. 2022). Examples of heuristic data augmentation techniques include flipping, distorting, and rotating images (Shorten and Khoshgoftaar 2019), using lexical substitution, inserting words, and shuffling sentences in texts (Kashefi and Hwa 2020), and deleting nodes and dropping edges in graphs (Zhao et al. 2021, 2023). While human knowledge can be used to improve data diversity and reduce over-fitting in heuristic methods, it is difficult to use a single heuristic method to preserve the different labels for different tasks (Balestriero et al. 2022; Cubuk et al., 2019). So, automated augmentation (Cubuk et al., 2019) learned from data to search for the best policy to combine a bunch of predefined heuristic augmentations. Generation models (Antoniou et al., 2017; Bowles et al., 2018; Han et al., 2022) create in-class examples. Other learning ideas such as FATTEN (Liu et al., 2018) and GREA (Liu et al., 2022a) learned to split the latent space for data augmentation. However, learning and augmenting from insufficient labels at the same time may limit the diversity of new examples and cause over-fitting. DCT leverages unlabeled data to avoid them.

Relationship between Data-Centric Approaches  As presented in Figure 7, perturb edges, delete nodes and mask attributes (Rong et al., 2019; Trivedi et al., 2022) for graphs are some heuristic ways for data augmentation. The augmented knowledge from them is mainly controlled by human prior knowledge on the perturbations and it often fails to be close to the task, i.e., , random perturbations hardly preserve labels for the augmented graphs. The learning to augment approaches learn from labeled graphs to perturb graph structures (Luo et al., 2022), to estimate graphons for different classes (Han et al., 2022), or to split the latent space for augmentation (Liu et al., 2022a). Although these approaches could preserve labels for the augmented graphs, they introduce less extra knowledge to improve the model prediction. In summary, graph data augmentation is effective in expanding knowledge for limited labels, but it makes no use of unlabeled graphs. Besides, the diversity and richness of the domain knowledge from augmented graphs are far from that contained in a large number of unlabeled graphs. To learn from unlabeled graphs, data-centric approaches like the self-training is assumed to be useful when the unlabeled and labeled data are from the same source. It is less studied when we have a single unified unlabeled source for different tasks.

B Additional Method Details

B.1 Upper bounding the mutual information

In Eq. (6), we use a leave-one-out variant of InfoNCE ($I_{\text{bound}}$) to derive the upper bound of mutual information. We summarize the derivation (Poole et al., 2019) here.

$$I_1(G'; G) = \mathbb{E}_{p(G,G')} \left[ \log \frac{p(G'|G)}{p(G')} \right]$$

$$= \mathbb{E}_{p(G,G')} \left[ \log \frac{p(G'|G)q(G')}{q(G')p(G')} \right]$$

$$= \mathbb{E}_{p(G,G')} \left[ \log \frac{p(G'|G)}{q(G')} \right] - \text{KL}(p(G')||q(G'))$$

$$\leq \mathbb{E}_{p(G,G')} \left[ \log \frac{p(G'|G)}{q(G')} \right]$$

The intractable upper bound is minimized when the variational approximation $q(G')$ matches the true marginal $p(G')$ (Poole et al., 2019). For each $G_i$, its augmented output $G_i'$, and $M - 1$ negative
examples with different labels, we could approximate \( q(G_i') = \frac{1}{M-1} \sum_{j \neq i} p(G_i'|G_j) \). So, we have

\[
\mathcal{I}_1(G_i', G_i) \leq \log \frac{p(G_i'|G_i)}{\frac{1}{M-1} \sum_{j=1,j \neq i}^M p(G_i'|G_j)} = \log \frac{p(G_i'|G_i)}{\sum_{j=1,j \neq i}^M p(G_i'|G_j)} + \log(M-1)
\]

(10)

### B.2 Extraction of Statistical Features on Graphs

For each molecule and polymer graph, we concatenate the following vectors or values for statistical feature extraction.

- the sum of the degree in the graph;
- the vector indicating the distribution of atom types;
- the vector containing the maximum, minimum and mean values of atoms weights in a molecule or polymer;
- the vector containing the maximum, minimum, and mean values of bond valence.

For each protein-protein interaction ego-graph in the biology field, we use the sorted vector of node degree distribution in the graph as the statistical features.

### B.3 Technical Details for Graph Data Augmentation with Diffusion Model

#### The Lookup Table from Atom Type to Node Embedding Space

Given a graph \( G \), we assume the node feature matrix on the graph is \( X \in \mathbb{R}^{n \times F_n} \), where \( n \) is the number of nodes. The edge feature matrix is \( E \in \mathbb{R}^{m \times F_e} \), where \( m \) is the number of edges. There are two ways for \( G \) to represent the graph structure in practice. We can use either the dense adjacency matrix \( A \in \mathbb{R}^{n \times n} \) or sparse edge index \( I_e \in \mathbb{R}^{2 \times m} \). The diffusion model (Jo et al. 2022) on graphs prefers the former, which is more straightforward for graph generations. The prediction model prefers the latter because of its flexibility, and less computational cost and time. The transformation between two types of graph structure representation takes additional time. Particularly for molecular graphs, the node features used for generation (one-hot encoding of the atom type) and for prediction (see the official package of OGBG for details) are different, which introduces extra time to process the graph data. For details, we (1) first need to extract discrete node attributes given the atom type and its neighborhoods; (2) we then need to use an embedding table to embed node attributes in a continuous embedding space; (3) the embedding features of nodes with their graph structure are inputted into the graph neural networks to get the latent representation for nodes. The reverse process for data augmentation in DCT may need to repeatedly process graph data with steps (1) and (2). It introduces additional time. To address these technical problems, we build up a lookup table to directly map the atom type to the node embedding. We average the node attributes for the same type of node within the batch. We then use the continuous node attributes as weights to average the corresponding node embedding according to the table.

#### Instantiations of SDE on Graphs

According to Song et al. (2021), we use the Variance Exploding (VE) SDE for the diffusion process. Given the minimal noise \( \sigma_{\text{min}} \) and the maximal noise \( \sigma_{\text{max}} \), the VE SDE is:

\[
dG = \sigma_{\text{min}} \left( \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}} \right)^t \sqrt{2 \log \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}} dw, \quad t \in (0, 1]
\]

(11)

The perturbation kernel is derived (Song et al. 2021) as:

\[
p_{\text{ot}}(G(t) \mid G(0)) = \mathcal{N} \left( G(t); G(0), \sigma_{\text{min}}^2 \left( \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}} \right)^{2t} \mathbf{I} \right), \quad t \in (0, 1]
\]

(12)

On graphs, we follow Jo et al. (2022) to separate the perturbation of adjacency matrix and node features:

\[
p_{\text{ot}}(G(t) \mid G(0)) = p_{\text{ot}}(A(t) \mid A(0))p_{\text{ot}}(X(t) \mid X(0)).
\]

(13)

https://github.com/snap-stanford/ogb/blob/master/ogb/utils/features.py
We perform experiments on 15 datasets, including eight classification and seven regression tasks from Algorithm 2 Diffusion-Based Graph Augmentation with PC Sampling

**Algorithm 1** Diffusion-Based Graph Augmentation with PC Sampling

**Input:** Graph $G$ with node feature $X$ and adjacency matrix $A$, the denoising function for node feature $s_X$ and adjacency matrix $s_A$, the fine-tune loss $L_{aug}$, Largoian MCMC step size $\beta$, scaling coefficient $\epsilon_1$

\[
\begin{align*}
A^{(0)} &\leftarrow A + z_A; \quad z_A \sim \mathcal{N}(0, I) \\
X^{(0)} &\leftarrow X + z_X; \quad z_X \sim \mathcal{N}(0, I) \\
\text{for } t = D - 1 \text{ to } 0 \text{ do} \\
&G^{(t+1)} \sim p_{t+1}(G^{(t+1)} | G^{(t+1)}) \quad \text{[inner-loop sampling with another PC sampler]} \\
&S_A = \frac{1}{2} s_A(G^{(t+1)}, t + 1) - \frac{1}{2} \alpha \nabla A(t) L_{aug}(G^{(t+1)}) \\
&S_X = \frac{1}{2} s_X(G^{(t+1)}, t + 1) - \frac{1}{2} \alpha \nabla X(t) L_{aug}(G^{(t+1)}) \\
&A^{(t)} \leftarrow A^{(t+1)} + g(t)S_A + g(t)z_A; \quad z_A \sim \mathcal{N}(0, I) \quad \text{[Predictor for adjacency matrix]} \\
&X^{(t)} \leftarrow X^{(t+1)} + g(t)S_X + g(t)z_X; \quad z_X \sim \mathcal{N}(0, I) \quad \text{[Predictor for node features]} \\
&A^{(t)} \leftarrow A^{(t)} + \frac{1}{2} S_A + \epsilon_1 \sqrt{\partial^2 z_A}; \quad z_A \sim \mathcal{N}(0, I) \quad \text{[Corrector for adjacency matrix]} \\
&X^{(t)} \leftarrow X^{(t)} + \frac{1}{2} S_X + \epsilon_1 \sqrt{\partial^2 z_X}; \quad z_X \sim \mathcal{N}(0, I) \quad \text{[Corrector for node features]} \\
\end{align*}
\]

end for

return $G' = (A^{(0)}, X^{(0)})$

The Algorithm of the Framework The algorithm of the proposed data-centric knowledge transfer framework is presented in Algorithm 2 and Algorithm 3. In detail, Algorithm 2 corresponds to Section 3.2 and Algorithm 3 corresponds to Section 3.3

**Algorithm 2** The Data-Centric Knowledge Transfer Framework: Learning from Unlabeled Graphs Framework: Generating Task-specific Labeled Graphs

**Input:** Given unlabeled graphs from the space $G^{[U]}$, randomly initialized score models $s_X$ and $s_A$ for node feature and graph adjacency matrix, respectively, the total diffusion time step $T$.

while $s_X$ and $s_A$ not converged do

Sample $G \in G^{[U]}$

Sample $t \in \text{Uniform}(1, 2, \ldots, T)$

Sample $z_A \sim \mathcal{N}(0, I)$

Sample $z_X \sim \mathcal{N}(0, I)$

Sample $G$ with $t$, $z_A$, $z_X$ and Eq. (13)

Optimize $s_A$ with the gradient:

\[
\nabla [z_A - s_A(G, t)]^2
\]

Optimize $s_X$ with the gradient:

\[
\nabla [z_X - s_X(G, t)]^2
\]

end while

**Algorithm 3** The Data-Centric Knowledge Transfer Framework: Generating Task-specific Labeled Graphs

**Input:** Given task $k$ with the graph-label space $(G, Y)$, a randomly initialized prediction model $f_0$, the well-trained score model $s = (s_X, s_A)$, the training data set $\{G_i, y_i\}_{i=1}^{N_t}$, total training epoch $e$, the hyper-paramter $n$ for current epoch $e$, from 1 to $e$ do

Train $f_0$ on current training data $\{G_i, y_i\}_{i=1}^{N_t}$

if $e_i$ is divisible by the augmentation interval then

Select $n$ graph-label pairs with the lowest training loss from $\{G_i, y_i\}_{i=1}^{N_t}$

Get the augmented examples $\{G'_i, y'_i\}_{i=1}^{n}$ by Algorithm 1 with the selected examples

Update $\{G_i, y_i\}_{i=1}^{N_t}$ with $\{G'_i, y'_i\}_{i=1}^{n}$, e.g., add $\{G'_i, y'_i\}_{i=1}^{n}$ to $\{G_i, y_i\}_{i=1}^{N_t}$.

end if

end for

**C Additional Experiments Set-ups**

We perform experiments on 15 datasets, including eight classification and seven regression tasks from chemistry, material science, and biology. We use Area under the ROC curve (AUC) for classification performance and mean absolute error (MAE) for regression.

**C.1 Molecule Classification and Regression Tasks**

Seven molecule classification and three molecule regression tasks are from open graph benchmark [Hu et al., 2020]. They were originally collected by MoleculeNet [Wu et al., 2018] and used to predict molecule properties. They include (1) inhibition to HIV virus replication in ogbg-HIV, (2) toxicological properties of 617 types in ogbg-ToxCast, (3) toxicity measurements such as nuclear receptors and stress response in ogbg-Tox21, (4) blood–brain barrier permeability in ogbg-BBBP, (5) inhibition to human β-secretase 1 in ogbg-BACE, (6) FDA approval status or failed clinical trial in ogbg-ClinTox,
(7) having drug side effects of 27 system organ classes in ogbg-SIDER, (8) predicting the property of lipophilicity in ogbg-Lipo, (9) predicting the water solubility (log solubility in mols per litre) from chemical structures in ogbg-ESOL, (10) predicting the hydration free energy of molecules in water in ogbg-FreeSolv. For all molecule datasets, we use the scaffold splitting procedure as the open graph benchmark adopted (Hu et al., 2020). It attempts to separate structurally different molecules into different subsets, which provides a more realistic estimate of model performance in experiments (Wu et al., 2018).

C.2 Polymer Regression Tasks

Four polymer regression tasks include GlassTemp, MeltingTemp, ThermCond, and \( \text{O}_2 \) \( \text{Perm} \). They are used to predict different polymer properties such as glass transition temperature (°C), melting temperature (°C), thermal conductivity (W/mK) and oxygen permeability (Barrer). GlassTemp and MeltingTemp are collected from PolyInfo, which is the largest web-based polymer database (Otsuka et al., 2011). The ThermCond dataset is from molecular dynamics simulation and is an extension from the dataset used in (Ma et al., 2022). The \( \text{O}_2 \) \( \text{Perm} \) dataset is created from the Membrane Society of Australasia portal, consisting of a variety of gas permeability data (Thornton et al., 2012). Since a polymer is built from repeated units, researchers often use a single unit graph with polymerization points as polymer graphs to predict properties. Different from molecular graphs, two polymerization points are two special nodes (see “∗” in Figure 2), indicating the polymerization of monomers (Cormack and Elorza, 2004). For all the polymer tasks, we randomly split by 60%/10%/30% for training, validation, and test.

C.3 Protein Classification Task

An additional task is protein function prediction using protein-protein interaction graphs (Hu et al., 2019). A node is a protein without attributes, an edge is a relation type between two proteins such as co-expression and co-occurrence. In our DCT, we treat all the relations as the undirected edge without attributes.

C.4 Baselines and Implementation

When implementing GIN (Xu et al., 2019), we tune its hyper-parameters for different tasks with an early stop on the validation set. We generally implement pre-training baselines following their own setting. For molecule and polymer property prediction and protein function prediction, the pre-trained GIN models with self-supervised tasks such as EDGE\text{PRED}, ATTR\text{MASK}, CONTEXT\text{PRED} in (Hu et al., 2019), INFO\text{MAX} (Velickovic et al., 2019) are available. So we directly use them. For other self-supervised methods, we implement their codes with default hyper-parameters. Following their settings, we use 2M ZINC15 (Sterling and Irwin, 2015) to pre-train GIN models for molecule and polymer property prediction. We use 306K unlabeled protein-protein interaction ego-networks (Hu et al., 2019) to pre-train the GIN for the downstream protein function property prediction. For self-training with real unlabeled graphs and INFO\text{GRAPH} (Sun et al., 2020), we use 113K QM9 (Ramakrishnan et al., 2014). For self-training with generated unlabeled graphs, we train the diffusion model (Jo et al., 2022) on the real QM9 dataset and then produce the same number of generated unlabeled graphs. To train the diffusion model in our DCT, we also use QM9 (Ramakrishnan et al., 2014).

D Additional Experiment Analysis

D.1 The Power of Diffusion Model to Learn from Unlabeled Graphs

In Table 3, when we replace the 133K QM9 with the 249K ZINC (Jo et al., 2022) to train the diffusion model, which nearly doubles the size of the unlabeled graphs and includes more atom types, we do not observe any additional improvement, and in some cases, even worse performance. It is possible because of the constraint of the current diffusion model’s capacity to model the data distribution for a much larger number of more complex graphs. It encourages powerful generative models in the future, which could be directly used to benefit predictive models under the proposed framework.
D.2 Chemical Validity of the Generated Graphs in Downstream Tasks

In Figure 6, we show through some examples that concepts, such as certain chemical rules from the unlabeled graphs, are successfully transferred to downstream tasks. To further validate this point, we gathered 1,000 task-specific graphs generated in the intermediate steps on the tasks of ogbg-BACE, ogbg-BBBP, ogbg-FreeSolv, and O2Perm. We then assessed the chemical validity of these graphs and observed that the validity is 92.8%, 87.9%, 97.4%, and 62.1%, respectively. Results show that transferring knowledge from pre-trained molecular data to target molecules yields relatively high chemical validity. However, the validity drops to 62% when transferring knowledge from pre-trained molecular data to target polymer data. This finding indicates that the transferability of chemical rules becomes more challenging when the distribution gap between the pre-training data and downstream task data is larger.