Molecular Data Programming: Towards Molecule Pseudo-labeling with Systematic Weak Supervision

Xin Juan[1], Kaixiong Zhou[2], Ninghao Liu[3], Tianlong Chen[4], Xin Wang[1,*]
[1]School of Artificial Intelligence, Jilin University, China
[2]Institute for Medical Engineering & Science, Massachusetts Institute of Technology, USA
[3]School of Computing, University of Georgia, USA
[4]CSAIL, Massachusetts Institute of Technology, USA
junxin22@mails.jlu.edu.cn, kz34@mit.edu,
ninghao.liu@uga.edu, tianlong@mit.edu,
xinwang@jlu.edu.cn

Abstract

The premise for the great advancement of molecular machine learning is dependent on a considerable amount of labeled data. In many real-world scenarios, the labeled molecules are limited in quantity or laborious to derive. Recent pseudo-labeling methods are usually designed based on a single domain knowledge, thereby failing to understand the comprehensive molecular configurations and limiting their adaptability to generalize across diverse biochemical context. To this end, we introduce an innovative paradigm for dealing with the molecule pseudo-labeling, named as Molecular Data Programming (MDP). In particular, we adopt systematic supervision sources via crafting multiple graph labeling functions, which covers various molecular structural knowledge of graph kernels, molecular fingerprints, and topological features. Each of them creates an uncertain and biased labels for the unlabeled molecules. To address the decision conflicts among the diverse pseudo-labels, we design a label synchronizer to differentiably model confidences and correlations between the labeling functions, which yields probabilistic molecular labels to adapt for specific applications. These probabilistic molecular labels are used to train a molecular classifier for improving its generalization capability. On eight benchmark datasets, we empirically demonstrate the effectiveness of MDP on the weakly supervised molecule classification tasks, achieving an average improvement of 9.5%*. The code is in: https://github.com/xinjuan1/MDP/.*

1. Introduction

Molecular machine learning has achieved remarkable success in chemical and biological applications, e.g., chemical property prediction [6, 22, 48, 51] and quantum chemistry calculations [25, 26, 34, 61], due to its generalization capability to differentiably learn molecule features and estimate the interested properties. The development processes have been catalyzed by the release of large labeled training datasets [23]. Unfortunately, manually collecting and labeling a sufficient amount of training data is often time-consuming and labor-intensive. This problem is exacerbated for molecular labeling from scientific domains, which rely on expensive wet-lab measurements or graph structure computations.

To enable the molecular machine learning with limited amount of labeled data, weakly supervised learning based on pseudo-label paradigm augments the training dataset with noisy sources without access to ground truth. These methods which can be roughly classified into two groups, i.e., self-training and graph kernel methods. The former [28, 56] repeats the process of training model on the available labeled data and yielding confident pseudo-labels to create a larger dataset. On the other hand, the later defines similarity measures for every pairwise graphs using kernel functions, including random-walk, Weisfeiler-Lehman, and graphlet kernels [11, 42, 43], where the similarity matrix among labeled and unlabeled molecules is leverage to infer pseudo-labels. However, while the self-training method is prone to over-confident predictions especially if the initial model is not accurate enough, the graph kernel algorithm is sensitive to the choice of an appropriate kernel function and cannot generalize well across different datasets.

To address the labeling overfitting and enhance generalization, data programming is proposed to unify multiple weak supervision sources and provide a more flexible and expressive pseudo-labeling [4, 49]. In particular, users define domain heuristics/rules in the form of labeling func-

[*]Corresponding author

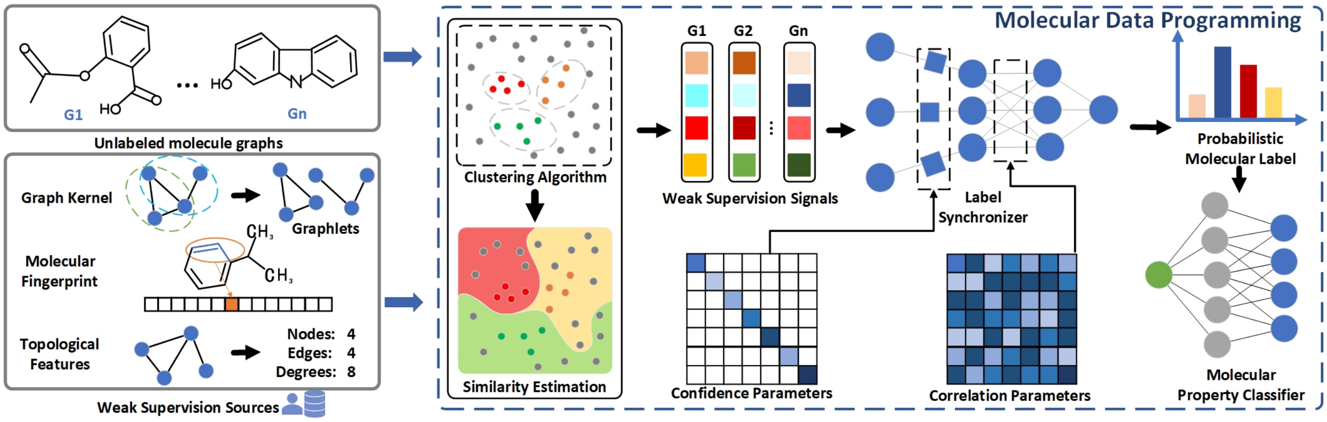

Figure 1. Diagram of MDP. (1) Designing LFs by expressing various weak supervision sources such as graph kernel, molecular fingerprint and topological features. (2) Employing the LFs over unlabeled data and learning a label synchronizer to aggregate the LF's outputs into probabilistic molecular labels. (3) Utilizing the generated probabilistic molecular labels, MDP trains a molecular property classifier to ensure that its predictions align with the pseudo-labels or ground-truth labels.

tions (LFs). Each LF generates cheap labels via capturing specific aspects of domain knowledge, and the multiple labels from diverse supervision sources are then combined to estimate the posterior probabilities of true labels for the unlabeled data. Synergizingly applying LFs can take into account the reliability and correlations of different LFs to create robust pseudo-labels. For example, Snorkel [36] is the first end-to-end system for leveraging LFs to serve as the programming interface, which generates weak supervision without any labeled training data. Snorkel MeTaL [37] extends to handle multi-task learning by exploring a hierarchical labeling model. Nevertheless, the aforementioned approaches are mainly adopted for text and image data.

However, the extension of programmatic weak supervision paradigm to pseudo-label molecules has to solve two key challenges. First, the canonical LFs are beingless to capture the nuances of biological or chemical properties. Unlike the text data whose label can be heuristically determined via named entities, molecules are often characterized by complex and intricate structures. The diversity of spatial structures in molecular datasets is immense, while the underlying biochemical mechanisms and interactions may not be fully understood. This lack of structural understanding makes it difficult to design rules that automatically capture the relevant features. Second, there may be decision conflicts among the different supervision sources. The simple majority vote may neutralize pseudo-labels capturing the desired properties for the specific applications.

To address the challenges mentioned above, as shown in Figure 1, we propose a novel molecular data programming (MDP) framework towards molecule pseudo-labeling with systematic weak supervisions. The key idea is to create robust pseudo-labels for unlabeled molecules by integrating knowledge from diverse domains and differentiably learning to adapt to specific biochemical environment. Particularly, we first define a set of molecular labeling functions to measure the structural similarities between molecules, including LFs based on graph kernels, molecular fingerprints, and topological features, to cover a wide range of biochemical context. Based on each similarity matrix consisted of all the molecule pairs, a clustering algorithm is applied to select a small subset of prominent examples from the multiple group centers. These prominent examples are labeled manually to serve for dual purposes: optimizing downstream classifiers and assigning weak supervision signals to the unlabeled molecules. We thus generate a large but noisy set of pseudo-labels according to the similarities between prominent examples and other unlabeled data. Furthermore, we design a label synchronizer to harmonize the decision conflicts between the different molecular labeling functions. The label synchronizer can dynamically and adaptively estimate the correlations between LFs to generate probabilistic molecular labels, which are leveraged to train a molecular property classifier to obtain the satisfactory generalization performance. The main contributions can be summarized as follows:

- We propose a novel molecular data programming framework for molecule pseudo-labeling, which incorporates multiple molecular labeling functions combing various graph domain knowledge to automatically generate a large amount of reasonable weak supervision signals for unlabeled molecules.
- We design a label synchronizer that dynamically adjusts the confidence of different molecular labeling functions and adaptively estimates the correlations between them, thereby mitigating uncertainty and generating the probabilistic molecular labels.

- Extensive experiments on benchmark datasets from the chemistry and biology domains demonstrate the effectiveness of our proposed framework for graph classification tasks when labeled molecules are scarce, where the average improvement is up to 9.5%.

2. Related Work

2.1. Molecular Representation Learning

Extended Connectivity Fingerprints (ECFP) [41] as the fixed representations are generated based on heuristics algorithms, aiming to maximize the information captured by the resulting feature vectors [5]. While these methods can undoubtedly be successful, they inevitably involve a trade-off by emphasizing specific molecular features while ignoring others. Recently, Simplified Molecular-Input Line-Entry System (SMILES) [52] is widely leveraged to denote molecular data, serving as the foundation for deriving a graph structure where heavy atoms are depicted as nodes and covalent bonds as edges. GNNs [14, 24, 44, 46, 60] have proven to be an effective tool due to their potential for processing graph-structured data. MPNNs [13] reformulate existing models into the Message Passing Neural Networks and explore novel variations. DGN [3] enables the definition of graph convolutions according based on topologically-derived directional flows by introducing the globally consistent anisotropic kernels. However, most existing works have not thoroughly explored the comprehensive integration of multiple domain knowledge sources into the molecular representation process, thus limiting their expressiveness.

2.2. Weakly Supervised Learning

Weakly supervised learning provides a framework for creating and combining weak labelers to derive pseudo-labels of unlabeled data [9, 17, 33, 38]. Data programming [39] is a new paradigm for the programmatic creation of training sets, in which users express the weak supervision approach as a labeling function. MeTaL [38] proposes a multi-task weak supervision setting, which views multiple weak supervision sources as different related sub-tasks of a problem. Recent advances in weakly supervised learning enhances the quality of pseudo-labels via improving the aggregation scheme or the weak labelers. For example, CLL [1] utilizes expected error to define a constrained space, containing the possible labels of the weak supervision signals. DP-SSL [57] generates probabilistic labels via combining data programming and multiple choice learning when there are only a few labeled samples per class. Inspired by data programming, our work yields a substantial number of reliable probabilistic labels by extracting a robust source of supervision signals from unlabeled data, thereby addressing the issue of label scarcity.

3. Preliminaries

3.1. Molecular Graph Embedding

Without loss of generalizability, we use graph neural networks (GNNs) to introduce the representation learning process. Given a molecular graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X_{\mathcal{V}}, X_{\mathcal{E}})$, where $\mathcal{V}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denote the set of nodes and edges, respectively. $X_{\mathcal{V}} = \{x_v | v \in \mathcal{V}\}$ and $X_{\mathcal{E}} = \{x_{uv} | (u, v) \in \mathcal{E}\}$ represent the features of nodes and edges, respectively. We leverage a GNN to learn embedding $h_v \in \mathbb{R}^d$ for each node $v \in \mathcal{V}$ and then obtain molecular graph embedding $h_g \in \mathbb{R}^d$ for the entire graph $\mathcal{G}$ after aggregating node embeddings. Specifically, for a L-layer GNN, the neighborhood aggregation process at the l-th layer can be defined as:

$$h_v^{(l)} = \mathrm{COM}^{(l)}(h_v^{(l-1)}, \mathrm{AGG}^{(l)}(\{(h_v^{(l-1)}, h_u^{(l-1)}, x_{uv}) \\ : u \in \mathcal{N}(v)\})), \tag{1}$$

where $h_v^{(l)}$ denotes the embedding of node v at the l-th layer, and $h_v^{(0)}$ is initialized as x_v. $\mathcal{N}(v)$ is the neighbor set of node v, and $\mathrm{COM}^{(l)}(\cdot)$ and $\mathrm{AGG}^{(l)}(\cdot)$ are combination and aggregation functions at the l-th layer. After propagating through L layers, the entire graph embedding of $\mathcal{G}$ can be derived through:

$$h_{\mathcal{G}} = \mathrm{READOUT}(\{h_v | v \in \mathcal{V}\}), \tag{2}$$

where $\mathrm{READOUT}(\cdot)$ is a permutation-invariant readout function, such as mean or summation pooling.

3.2. Pseudo-label Generation

The goal of pseudo-label generation is to infer the unknown label vectors of molecules without access to their ground-truths. Particularly, considering a molecular dataset $\mathcal{D} = \{\mathcal{G}_1, \cdots, \mathcal{G}_N\}$, the pseudo-label generation defines a labeling function to annotate the unlabeled molecules according to prior biochemical knowledge: $\mathcal{F}(\mathcal{G}_i) \in \{\mathcal{Y} \cup \{\phi\}\}$, where $\mathcal{F}$ denotes a specific labeling function, $\mathcal{Y}$ denotes the molecular class space of a downstream application, and ϕ means null label if the molecule's class cannot be confidently generated. Let $\widetilde{Y} = \{\mathcal{F}(\mathcal{G}_i)\}_{i=1}^{N}$ denote the set of inferred pseudo-labels, which will be adopted to train the molecular graph embedding learning introduced above.

4. The Proposed Framework

In this section, we introduce MDP via solving the two challenges mentioned in Section 1. The key idea of MDP is to generate robust pseudo-labels for unlabeled molecules by integrating knowledge from diverse domains and synergizingly applying different labeling functions. The workflow of MDP unfolds in three main stages.

1 **Designing Molecular Labeling Functions:** Rather than hand-crafting labels for training data, MDP leverages dif-

ferent LFs to yield weak supervision signals for unlabeled molecules. These LFs are crafted by considering various molecular structural knowledge of graph kernels, molecular fingerprints, and topological features.

2 **Estimating Correlations between Labeling Functions:** Given the weak supervision signals from the diverse domains, MDP leverages a label synchronizer to generate probabilistic molecular labels by estimating correlations between the different LFs and combining their biased pseudo-labels. The learning process of the label synchronizer is automatic and dataset-agnostic.

3 **Training the Molecular Property Classifier:** The output of label synchronizer for the unlabeled molecules is used to train a molecular property classifier, such as graph neural networks. By leveraging a large amount of unlabeled molecules, the molecular property classifier can improve its coverage and robustness on unseen data.

4.1. Designing Molecular Labeling Functions

In the traditional molecular feature engineering, we typically resort to manual creation of features to learn low-dimensional representations of molecules. For instance, researchers have developed hand-crafting molecular descriptors such as atom types, bond types, topological features, or quantum chemical properties to capture molecular characteristics [22, 58]. These engineered features play a crucial role in understanding the inherent correlation between molecular structure and its properties. Nonetheless, much of the existing work focuses on a limited set of rules for creating molecule features, thereby constraining their adaptability and generalization in a wide range of biochemical application context. In this work, we define the molecular labeling function to harness the potential of weakly supervised learning based on the data programming paradigm. These molecular labeling functions offer a more adaptable and data-driven approach to create weak supervision signals that can capture a broader range of structural and property characteristics from diverse perspectives.

Definition 1 (Molecular labeling function). The molecular labeling function, denoted as $\mathcal{F}$, infers the pseudo-labels for molecules. Specifically, each LF consists of two modules: similarity estimation and weak supervision signals creation. The similarity estimation measures the similarity score between every pair of molecules using a quantitative function $\mathcal{S}$: $\mathbf{S}_{ij} = \mathcal{S}(\mathcal{G}_i, \mathcal{G}_j)$, where $\mathbf{S}_{ij}$ is the (i, j)-th element of $\mathbf{S}$ corresponding to molecules $\mathcal{G}_i$ and $\mathcal{G}_j$. Based on matrix $\mathbf{S}$, the pseudo-labels are derived via maximizing the discrepancy of the different classes.

Similarity estimation. Due to the intricate structural complexity of molecular data and the unknown atomic interaction mechanisms that determine molecular properties, it's unfeasible to directly generate weak supervision signals through heuristic rules, or external knowledge bases, without any manual supervision. To address this challenge, inspired by metric learning [27], we assess the similarity between molecules before deriving weak supervision signals for unlabeled molecular graphs.

Given the molecular dataset $\mathcal{D}$, we utilize a number of z similarity estimation methods to derive multiple similarity matrices $\{\mathbf{S}^1, \cdots, \mathbf{S}^z\}$ by measuring the similarity between any two molecular graphs. $\mathbf{S}^k$ is a $N \times N$ matrix. We design LFs from three perspectives:

1. Graph-Kernel based LFs: Computing similarity by comparing occurrences of graphlets [43], shortest-paths [42] or subtrees [42] across multiple graphs.
2. Molecular-fingerprint based LFs: Computing similarity according to the tanimoto coefficience by using various encoding methods such as MACCS [8] and Morgan [30] encoding.
3. Topological-feature based LFs: Computing similarity based on differences in topological features, such as the number of nodes, degrees and so on.

After obtaining the set of similarity matrices, a clustering algorithm is employed on each of them to choose a small subset of prominent examples from various group centers. These selected examples are then manually labeled to form the labeled set $\mathcal{D}_L = \{\mathcal{G}_1^L, \cdots, \mathcal{G}_{|\mathcal{D}_L|}^L\}$. This process resembles active learning but does not necessitate any dataset-specific parameter learning. Here $\mathcal{D}_L$ and $\{\mathbf{S}^1, \cdots, \mathbf{S}^z\}$ are leveraged in the following subsection to yield weak supervision signals for unlabeled molecular graphs.

Weak supervision signals creation. To unify the multi-class molecular classification tasks containing a number of m labels, we decompose them into m subtasks. Each subtasks is to classify whether the molecule contains a specific property ($c = 1$) or not ($c = -1$). For each similarity matrix $\mathbf{S}^k$, we derive the pseudo-labels for unlabeled molecular graphs according to the closeness between them with the labeled molecular graphs from $\mathcal{D}_L$. Specifically, the process of deriving the weak supervision signal for the unlabeled graph $\mathcal{G}_j^U$ can be formulated as:

$$\Lambda_{j,m}^k = \begin{cases} \max_c(\frac{1}{Z_m} exp\{\sum_{i=1}^{|\mathcal{D}^L|} \mathbb{I}\left(y_{\mathcal{G}_i^L,m} = c\right) * \mathcal{W}_{ij}^k\}), \text{if } c = 1 \\ -\max_c(\frac{1}{Z_m} exp\{\sum_{i=1}^{|\mathcal{D}^L|} \mathbb{I}\left(y_{\mathcal{G}_i^L,m} = c\right) * \mathcal{W}_{ij}^k\}), \text{if } c = -1, \end{cases} \tag{3}$$

where $\Lambda_{j,m}^k$ represents a scalar called weak supervision signal for the unlabeled graph $\mathcal{G}_j^U$ generated by the k-th molecular labeling function on the m-th binary prediction task. $\mathcal{W}_{ij}^k$ is the weight computed according to the the normalized similarity between the labeled graph $\mathcal{G}_i^L$ and the unlabeled graph $\mathcal{G}_j^U$, which can be formulated as:

$$\mathcal{W}_{ij}^k = exp\Big\{ - \frac{\text{Normalize}\left(\text{S}_{ij}^k\right)}{\gamma}\Big\}, \tag{4}$$

The goal of normalizing the similarity matrix is to mitigate the impact of the different magnitude of the similarity matrices. Z_m is the normalization factor, which is defined as:

$$\begin{aligned} Z_m = exp\{&\sum_{i=1}^{|\mathcal{D}^L|} \mathbb{I}\left(y_{\mathcal{G}_i^L,m} = -1\right) * \mathcal{W}_{ij}^k\} \\ &+ exp\{\sum_{i=1}^{|\mathcal{D}^L|} \mathbb{I}\left(y_{\mathcal{G}_i^L,m} = 1\right) * \mathcal{W}_{ij}^k\}. \end{aligned} \tag{5}$$

4.2. Estimating Correlations between Molecular Labeling Functions

Each labeling function is a noisy voter and may contains conflicts with others. Thus, when integrating knowledge from diverse domains to establish molecular labeling functions, an effective application of weak supervision paradigm faces two primary challenges. First, how to enhance the accuracy of generated pseudo-labels by resolving the conflicts between various molecular labeling functions. Second, how to transmit the critical information about the quality of pseudo-labels to the molecular property classifier under training. To address these two challenges, we propose a label synchronizer G_ϵ to provide approximate indicators of the target classification for the unlabeled data according to the multiple weak supervision signals.

We first initialize a learnable vector Ω from a uniform distribution $\Omega \sim U\left(0,1\right)^z$ to endow the label synchronizer with the capability of identifying the confidence of different labeling functions. Next, we aim for the label synchronizer G_ϵ to effectively solve the issue of potentially overlapping and conflict sources. According to Kolmogorov's theorem [7], multilayer perceptron (MLP) with only one single hidden layer is capable of approximating various continuous functions. Furthermore, the simplicity of MLP is crucial for establishing the label synchronizer when a few labeled molecular data is available. Therefore, we equip the label synchronizer with MLP to capture the intricate correlations among diverse molecular labeling functions. $\epsilon = \{\Omega \in \mathbb{R}^{z\times 1}, \text{W}^{\text{MLP}} \in \mathbb{R}^{z\times z}\}$.

Specifically, after obtaining the weak supervision signals matrices $\{\Lambda^1, \Lambda^2, \cdots, \Lambda^z\}$ across multiple graph labeling functions, we first utilize Ω to represent the confidence of each molecular labeling function. Then, for a unlabeled molecule $\mathcal{G}_j^U$, the generation process of probabilistic molecular label is formulated as:

$$\widetilde{y}_{j,m} = \sigma(\text{W}^{\text{MLP}}, \|\{\Omega_k \Lambda_{j,m}^k\}_{k=1}^z), \tag{6}$$

where $\|\{\Omega_k \Lambda_{j,m}^k\}_{k=1}^z$ represents concatenating the weak supervision signals of unlabeled graph $\mathcal{G}_j^U$ considering the confidence of all molecular labeling functions, σ represents the sigmoid function and $\widetilde{y}_{j,m}$ denotes the pseudo-label for unlabeled graph $\mathcal{G}_j^U$ on the m-th binary prediction task. By optimizing W^{MLP}, when labeling functions exhibit conflicting weak supervision signals for the majority of unlabeled data, the value of W^{MLP} is negative. Conversely, it becomes positive when they are in concordance. We use $\widetilde{\mathbf{Y}}$ as the probabilistic molecular label matrix to train a molecular property classifier in the subsequent section.

4.3. Training a Molecular Property Classifier

In MDP, the ultimate objective is to train a molecular property classifier capable of generalizing beyond the information provided by the labeling functions. Throughout the training process, we train the molecular property classifier, denoted as D_θ, on both the probabilistic labels $\widetilde{\mathbf{Y}}$ and the ground-truth labels $\mathbf{Y}$. This is achieved by minimizing two noise-aware variants of the binary cross-entropy loss: $\mathcal{L}(D_\theta(\mathcal{G}_i), y_i)$ and $\mathcal{L}(D_\theta(\mathcal{G}_j), \widetilde{y}_j)$. The optimization process of θ can be expressed as:

$$\begin{aligned} \theta^* = \underset{\theta^*}{\text{argmin}} \Bigg[&\sum_{i=1}^{|\mathcal{D}_L|} \mathbb{E}_{y_i\sim\mathbf{Y}}\left[\mathcal{L}(D_\theta(\mathcal{G}_i), y_i)\right] \\ &+ \lambda \sum_{j=1}^{|\mathcal{D}_U|} \mathbb{E}_{\widetilde{y}_j\sim\widetilde{\mathbf{Y}}}\left[\mathcal{L}(D_\theta(\mathcal{G}_j), \widetilde{y}_j)\right] \Bigg], \end{aligned} \tag{7}$$

where λ is a tuning parameter to control the emphasis between $\mathcal{L}(D_\theta(\mathcal{G}_i), y_i)$ and $\mathcal{L}(D_\theta(\mathcal{G}_j), \widetilde{y}_j)$. As for the label synchronizer G_ϵ, we construct a adaptive loss item to optimize the parameter ϵ for searching the optimal values of the parameters Ω^* and W_{MLP}^* by minimizing this adaptive loss function:

$$\epsilon^* = \underset{\epsilon^*}{\text{argmin}} \sum_{i=1}^{|\mathcal{D}_L|} \mathbb{E}_{y_i\sim\mathbf{Y}}\left[\mathcal{L}(G_\epsilon(\{\Lambda_i^1, \Lambda_i^2, \cdots, \Lambda_i^z\}), y_i)\right] \tag{8}$$

The total loss function of MDP consists of three binary cross entropy loss items, which can be written as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_s(\bar{\mathbf{Y}}, \mathbf{Y}) + \lambda \mathcal{L}_u(\bar{\mathbf{Y}}, \widetilde{\mathbf{Y}}) + \mathcal{L}_{\text{adap}}(\widetilde{\mathbf{Y}}, \mathbf{Y}) \\ &= \sum_{i=1}^{|\mathcal{D}_L|} \mathbb{E}_{y_i\sim\mathbf{Y}}\left[\mathcal{L}(D_\theta(\mathcal{G}_i), y_i)\right] + \lambda \sum_{j=1}^{|\mathcal{D}_U|} \mathbb{E}_{\widetilde{y}_j\sim\widetilde{\mathbf{Y}}}\left[\mathcal{L}(D_\theta(\mathcal{G}_j), \widetilde{y}_j)\right] \\ &+ \sum_{i=1}^{|\mathcal{D}_L|} \mathbb{E}_{y_i\sim\mathbf{Y}}\left[\mathcal{L}(G_\epsilon(\{\Lambda_i^1, \Lambda_i^2, \cdots, \Lambda_i^z\}), y_i)\right]), \end{aligned} \tag{9}$$

where $\mathcal{L}_{\text{s}}$ is the supervised loss over all labeled graphs and $\mathcal{L}_{\text{u}}$ denotes the unsupervised loss computed on all unlabeled graphs. $\bar{\mathbf{Y}}$ represents the prediction from the classification model D_θ. During the training process, we first learn ϵ with $\mathcal{L}_{adap}$. Then the pseudo-labels $\widetilde{\mathbf{Y}}$ are fixed and used to learn the classifier. We list the pseudo code in Appendix A.

Table 1. Test AUC performance of different methods on various molecular prediction benchmarks under the weak supervision setting. The best results are in bold and the second best results are underlined.

Method	BACE	BBBP	SIDER	ClinTox	Tox21	ToxCast	HIV	MUV
GCN	0.488±0.022	0.509±0.017	0.513±0.006	0.455±0.007	0.525±0.013	0.600±0.002	0.516±0.075	0.535±0.010
GAT	0.532±0.032	0.519±0.014	0.514±0.009	0.450±0.012	0.533±0.009	0.603±0.013	0.522±0.094	0.541±0.017
GraphSAGE	0.504±0.025	0.522±0.017	0.517±0.003	0.440±0.013	0.520±0.016	0.594±0.007	0.518±0.088	0.515±0.009
GIN	0.497±0.026	0.517±0.024	0.571±0.013	0.433±0.012	0.486±0.012	0.609±0.005	0.519±0.044	0.502±0.013
Pseudo-labeling	0.542±0.038	0.542±0.046	0.552±0.026	0.497±0.018	0.543±0.028	0.499±0.039	0.558±0.056	0.493±0.025
Self-training	0.532±0.045	0.527±0.038	0.520±0.027	0.502±0.023	0.556±0.015	0.503±0.042	0.536±0.043	0.524±0.033
infomax	0.524±0.031	0.462±0.043	0.563±0.004	0.460±0.023	0.510±0.016	0.610±0.004	0.515±0.044	0.573±0.016
context_pred	0.530±0.070	0.547±0.014	0.584±0.009	0.494±0.006	0.533±0.014	0.616±0.005	0.562±0.059	0.585±0.023
attr_mask	0.553±0.047	0.501±0.011	0.555±0.008	0.454±0.018	0.607±0.010	0.608±0.002	0.527±0.043	0.512±0.039
edge_pred	0.504±0.020	0.552±0.013	0.595±0.006	0.439±0.002	0.569±0.010	**0.617±0.003**	0.549±0.026	0.552±0.028
GraphLoG	0.586±0.016	0.527±0.012	0.524±0.033	0.494±0.056	0.555±0.011	0.550±0.002	0.503±0.052	0.560±0.049
perturb_edge	0.464±0.006	0.473±0.026	0.581±0.009	0.512±0.031	0.559±0.013	0.609±0.006	0.557±0.044	0.513±0.052
drop_node	0.545±0.021	0.532±0.041	0.558±0.004	0.545±0.057	0.523±0.016	0.598±0.006	0.575±0.059	0.539±0.017
subgraph	0.441±0.021	0.454±0.021	0.518±0.005	0.521±0.024	0.469±0.011	0.609±0.003	0.564±0.028	0.487±0.059
Ours	**0.632±0.023**	**0.618±0.011**	**0.613±0.018**	**0.615±0.043**	**0.653±0.006**	0.602±0.007	**0.665±0.019**	**0.604±0.022**

Table 2. The effect of different components on molecular property prediction tasks.

Confidence	*Correlation*	*Clustering*	BACE	BBBP	ClinTox	SIDER	Tox21
✓			0.484±0.038	0.565±0.032	0.523±0.053	0.513±0.043	0.589±0.020
	✓		0.546±0.038	0.572±0.038	0.543±0.060	0.506±0.016	0.588±0.016
		✓	0.617±0.033	0.580±0.029	0.552±0.037	0.581±0.015	0.608±0.009
✓	✓		0.529±0.096	0.568±0.009	0.500±0.022	0.527±0.027	0.580±0.020
✓		✓	0.602±0.023	0.566±0.027	0.588±0.047	0.553±0.013	0.619±0.003
	✓	✓	0.619±0.025	0.606±0.018	0.600±0.045	0.596±0.008	0.648±0.004
✓	✓	✓	0.632±0.023	0.618±0.011	0.615±0.043	0.613±0.018	0.653±0.006

5. Experiments

In this section, we conduct extensive experiments on benchmark datasets to evaluate the effectiveness of MDP.

5.1. Experimental Settings

▷ **Datasets**. Following the settings of previous molecular graph tasks, we evaluate our framework using 8 binary classification datasets: BBBP [29], Tox21 [12], ToxCast [40], SIDER [19], ClinTox [31], MUV [10], HIV [2], BACE [45] from MoleculeNet [53], a benchmark for molecular property prediction. Details are in Appendix B.

▷ **Baseline models**. To verify the effectiveness of MDP, we evaluate it by comparing it with four types of related methods: (1) *vanilla GNN* including GCN [18], GAT [50], GraphSAGE [15] and GIN [54]; (2) *semi-supervised learning methods* including self-training [21] and pseudo-labeling [20]; (3)*self-supervised predictive learning methods* including edge prediction (edge_pred) [16], attribute masking (attr_mask) [16], context prediction (context_pred) [16] and infomax [16]; (4) *self-supervised contrastive learning methods* including node dropping (drop_node) [59], edge perturbation (perturb_edge) [59], subgraph extraction (subgraph) [59] and GraphLoG [55].

▷ **Implementation**. The details of implementation are in Appendix C.

5.2. Evaluation on Molecular Graph Datasets

Q: Whether our proposed framework MDP can outperform the baselines on molecular datasets under weakly supervised learning? Yes, one key advantage of MDP is to infer high-quality pseudo-labels by combining diverse domain knowledge, which can help the molecular property classifier identify more features that co-occur with the heuristics encoded in the labeling functions.

▷ **Comparison between traditional semi-supervised learning methods and vanilla GNN models**. The comparison results are shown in Table 1, from which we make following observations. ❶ *The traditional semi-supervised learning methods (pseudo-labeling and self-training) can achieve better performance than the vanilla GNN in most cases.* The experimental results demonstrate that incorporating pseudo-labels learned from unlabeled data into training process is usefulness. ❷ *When the number of binary classification tasks increases, the efficacy of pseudo-labeling and self-training may diminish.* As the number

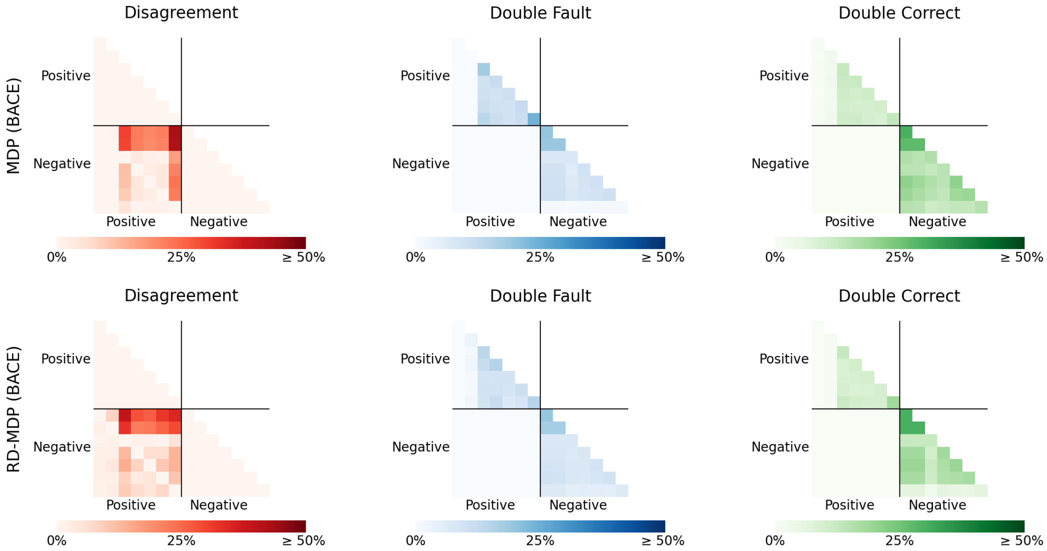

Figure 2. Diversity metrics among labeling functions on BACE: Disagreement (left), Double Fault (center), Double Correct (right). Each cell in the matrix denotes the coverage of training instances, marked by color intensity, where both molecular labeling functions i and j label the same example. Note that certain blocks are inherently zero. For instance, the Double Correct metric represents situations when two molecular labeling functions simultaneously provide the same correct label, resulting in a zero value in the Positive/Negative block.

of binary classification tasks increases, depending solely on a single domain of knowledge for pseudo-label generation may result in unreliable pseudo-labels, potentially leading to limited or even detrimental performance improvement.

▷ **Comparison between self-supervised learning methods and vanilla GNN models.** As shown in Table 1, we observe ❶ *the self-supervised learning methods can obtain better performance than the vanilla GNN.* Due to the additional supervision information extracted from unlabeled data, the experimental results illustrate that utilizing considerable unlabeled data can be effective to alleviate the bottleneck of labeled data. However, the pretext tasks of these methods are not specifically designed for downstream tasks. Therefore, there exists an inherent training objective gap between the pretext and downstream tasks.

▷ **Comparison between MDP and all baseline models.** From Tabel 1, we can easily observe that ❶ *MDP can outperform all baselines on all datasets, except ToxCast.* With proper denoising scheme and strong association of pseudo-labels, MDP can effectively resolve the conflicts and overlapping issue between labeling functions. Experimental results demonstrate the advantages of MDP over others by integrating multiple LFs and estimating correlations to generate pseudo-labels with high quality. However, due to the large number of binary classification tasks and label sparsity within the ToxCast dataset, MDP's performance is not satisfactory. While the simplicity of the label synchronizer is its advantage in weakly supervised learning, it encounters challenges in adapting effectively when the generated pseudo-labels become complex.

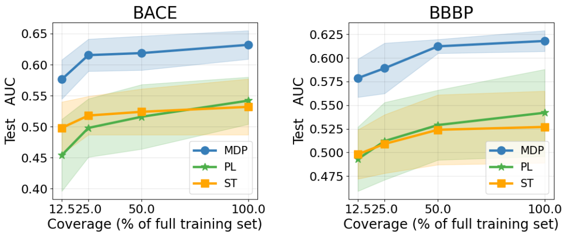

Figure 3. Plot of increasing labeling sample coverage (x-axis), v.s., accuracy (y-axis) using test curves.

5.3. Ablation Studies.

Q: How do different components of MDP contribute to the performance of weak supervision? We carry out an ablation study to analyze the effect of each component of MDP. We can easily find that ❶ *using only clustering to sample labeled graphs can be helpful for improving model's performance.* This phenomenon illustrates that calculating inter-instance similarity by aggregating multiple distance matrices can effectively sample significant labeled data without requiring human labeling efforts or additional parameter learning processes. ❷ *Combining clustering and estimating the correlation between molecular labeling functions (5th row) outperforms almost all baselines except the full model (last row).* This experiment emphasizes that resolving conflicts between various labeling functions significantly enhances the reliability of probabilistic molecular labels, ultimately improving the model's performance. ❸ *The full model (last row) achieves the best performance.* This experimental result demonstrates that estimating the correlation between molecular labeling functions and dynamically learning the confidence of different molecular labeling functions are complementary to each other.

5.4. Experiment with Fewer Labeled Examples.

Q: Whether MDP can perform well as the number of pseudo-labeled data changes? When we increase the coverage of labeling sample, as depicted in Figure 3, it becomes evident that MDP consistently outperforms pseudo-labeling (PL) and self-training (ST). Furthermore, the model trained using data generated with different types of labeling functions performs quite well with an AUC of 0.576 when only 12.5% of the full BACE training set is labeled. As the coverage of labeled samples grows, the performance demonstrates improvement. Notably, when we label the entire training set, the classifier attains its highest level of performance. The same trend is observed on the other datasets, as shown in Figure 6 in Appendix D.2.

5.5. Diversity Measures.

Q: Whether is it advantageous to select molecular labeled data using the clustering algorithm instead of random sampling? To enhance our understanding of the diversity among various labeling functions, we calculate metrics based on ensemble diversity measures to answer the question of whether using a clustering algorithm for sampling molecular labeled data is more beneficial than random sampling. In particular, we utilize a 2×2 contingency table T to count for the votes for all unlabeled examples, which encompasses samples covered by $T_{00}^{ij} + T_{10}^{ij} + T_{01}^{ij} + T_{11}^{ij}$ in binary classification. T^{ij} is the total number of label pairs emitted by labeling function i and j. We calculate the diversity using the following metrics, all of which are normalized with respect to the size of the unlabeled training set: Agreement:$= T_{00} + T_{11}$. Disagreement:$= T_{10} + T_{01}$. Double Fault:$= T_{00}$. Double Correct:$= T_{11}$.

Figure 2 presents a heatmap illustrating the pairwise diversity of BACE. In this heatmap, "RD-MDP" represents that MDP selects molecular labeled data using the random sampling. Notably, there is more variation (disagreement) and less agreement (double fault and double correct) in the RD-MDP when compared to the standard MDP method. Similar phenomenons are observed on the BBBP dataset, as shown in Figure 5 in Appendix D.1. This observation highlights that selecting labeled graphs with a clustering algorithm can achieve stronger correlation and less variation in weak supervision signals, which undoubtedly lead to greater ensemble efficiency.

5.6. Training and Validation Curves.

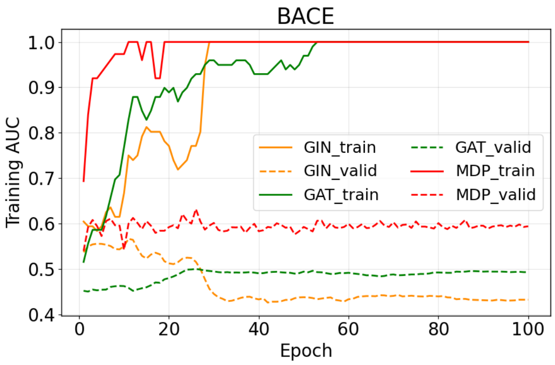

Figure 4. Training and validation curves of different models.

Q: Whether MDP achieves faster training and validation convergence than the vanilla GNN models? Beyond predictive performance improvement, we show the training and validation curves of different methods in molecular prediction tasks in Figure 4. We find that MDP achieves faster training convergence and higher validation performance than other vanilla GNN models on the dataset BACE. The training curve of MDP (the red solid line) takes about 20 epochs to reach convergence, while other vanilla GNN models takes about 30 epochs to gradually start converging. Meantime, the validation performance of MDP (the red dashed line) is the highest than other GNN models, which are rather significant. These experimental results illustrate the effectiveness of the MDP.

6. Conclusion

We develop a unifying framework for weakly supervised learning on molecular data. Crucial to the success of the proposed framework is to consider various domain knowledge to construct different molecular labeling functions and utilize a label synchronizer to estimate the correlations between them. This ensures that MDP can be effective on different datasets without additional any pre-training or fine-tuning. Extensive experiments on multiple benchmark datasets from the chemistry and biology domains and different GNN backbones demonstrate the effectiveness of our proposed framework for graph classification tasks under weak supervision.

Acknowledgement

This work was supported by a grant from the National Natural Science Foundation of China under grants (No.62372211, 62272191), and the Foundation of the National Key Research and Development of China (No.2021ZD0112500), and the International Science and Technology Cooperation Program of Jilin Province (No.20230402076GH, No.20240402067GH), and the Science and Technology Development Program of Jilin Province (No. 20220201153GX).

References

[1] Chidubem Arachie and Bert Huang. Constrained labeling for weakly supervised learning. In *Uncertainty in Artificial Intelligence*, pages 236–246. PMLR, 2021. 3

[2] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 6, 12

[3] Dominique Beaini, Saro Passaro, Vincent Létourneau, Will Hamilton, Gabriele Corso, and Pietro Liò. Directional graph networks. In *International Conference on Machine Learning*, pages 748–758. PMLR, 2021. 3

[4] Oishik Chatterjee, Ganesh Ramakrishnan, and Sunita Sarawagi. Data programming using continuous and quality-guided labeling functions. *CoRR*, abs/1911.09860, 2019. 1

[5] Jianwen Chen, Shuangjia Zheng, Ying Song, Jiahua Rao, and Yuedong Yang. Learning attributed graph representations with communicative message passing transformer. *arXiv preprint arXiv:2107.08773*, 2021. 3

[6] Yu-Neng Chuang, Guanchu Wang, Chia-Yuan Chang, Kwei-Herng Lai, Daochen Zha, Ruixiang Tang, Fan Yang, Alfredo Costilla Reyes, Kaixiong Zhou, Xiaoqian Jiang, et al. Discoverpath: A knowledge refinement and retrieval system for interdisciplinarity on biomedical research. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5021–5025, 2023. 1

[7] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989. 5

[8] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002. 4

[9] Daniel Fu, Mayee Chen, Frederic Sala, Sarah Hooper, Kayvon Fatahalian, and Christopher Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. In *International Conference on Machine Learning*, pages 3280–3291. PMLR, 2020. 3

[10] Eleanor J Gardiner, John D Holliday, Caroline O'Dowd, and Peter Willett. Effectiveness of 2d fingerprints for scaffold hopping. *Future medicinal chemistry*, 3(4):405–414, 2011. 6, 12

[11] Thomas Gärtner, Peter Flach, and Stefan Wrobel. On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 129–143. Springer, 2003. 1

[12] Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology*, 23(10):1294–1301, 2016. 6, 12

[13] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017. 3

[14] Kai Guo, Kaixiong Zhou, Xia Hu, Yu Li, Yi Chang, and Xin Wang. Orthogonal graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3996–4004, 2022. 3

[15] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017. 6

[16] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019. 6

[17] Zhimeng Jiang, Kaixiong Zhou, Zirui Liu, Li Li, Rui Chen, Soo-Hyun Choi, and Xia Hu. An information fusion approach to learning with instance-dependent label noise. In *International Conference on Learning Representations*, 2021. 3

[18] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 6

[19] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2016. 6, 12

[20] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 6

[21] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 6

[22] Zhen Li, Mingjian Jiang, Shuang Wang, and Shugang Zhang. Deep learning methods for molecular representation and property prediction. *Drug Discovery Today*, page 103373, 2022. 1, 4

[23] Zirui Liu, Haifeng Jin, Ting-Hsiang Wang, Kaixiong Zhou, and Xia Hu. Divaug: Plug-in automated data augmentation with explicit diversity maximization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4762–4770, 2021. 1

[24] Zirui Liu, Chen Shengyuan, Kaixiong Zhou, Daochen Zha, Xiao Huang, and Xia Hu. Rsc: accelerate graph neural networks training via randomized sparse computations. In *International Conference on Machine Learning*, pages 21951–21968. PMLR, 2023. 3

[25] Sönke Lorenz, Axel Groß, and Matthias Scheffler. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chemical Physics Letters*, 395(4-6):210–215, 2004. 1

[26] Sönke Lorenz, Matthias Scheffler, and Axel Gross. Descriptions of surface chemical reactions using a neural network representation of the potential-energy surface. *Physical Review B*, 73(11):115431, 2006. 1

[27] Guixiang Ma, Nesreen K Ahmed, Theodore L Willke, and Philip S Yu. Deep graph similarity learning: A survey. *Data Mining and Knowledge Discovery*, 35:688–725, 2021. 4

[28] Hehuan Ma, Feng Jiang, Yu Rong, Yuzhi Guo, and Junzhou Huang. Robust self-training strategy for various molecular biology prediction tasks. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–5, 2022. 1

[29] Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, 52(6):1686–1697, 2012. 6, 12

[30] Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965. 4

[31] Paul A Novick, Oscar F Ortiz, Jared Poelman, Amir Y Abdulhay, and Vijay S Pande. Sweetlead: an in silico database of approved drugs, regulated chemicals, and herbal isolates for computer-aided drug discovery. *PloS one*, 8(11):e79568, 2013. 6, 12

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 12

[33] Rattana Pukdee, Dylan Sam, Pradeep Kumar Ravikumar, and Nina Balcan. Label propagation with weak supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[34] LM Raff, M Malshe, M Hagan, DI Doughan, MG Rockley, and R Komanduri. Ab initio potential-energy surfaces for complex, multichannel systems using modified novelty sampling and feedforward neural networks. *The Journal of chemical physics*, 122(8), 2005. 1

[35] Bharath Ramsundar, Peter Eastman, Patrick Walters, and Vijay Pande. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. O'Reilly Media, 2019. 12

[36] Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason A. Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *VLDB J.*, 29(2-3):709–730, 2020. 2

[37] Alexander Ratner, Braden Hancock, Jared Dunnmon, Roger E. Goldman, and Christopher Ré. Snorkel metal: Weak supervision for multi-task learning. In Sebastian Schelter, Stephan Seufert, and Arun Kumar, editors, *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning, DEEM@SIGMOD 2018, Houston, TX, USA, June 15, 2018*, pages 3:1–3:4. ACM, 2018. 2

[38] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4763–4771, 2019. 3

[39] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29, 2016. 3

[40] Ann M Richard, Richard S Judson, Keith A Houck, Christopher M Grulke, Patra Volarath, Inthirany Thillainadarajah, Chihae Yang, James Rathman, Matthew T Martin, John F Wambaugh, et al. Toxcast chemical landscape: paving the road to 21st century toxicology. *Chemical research in toxicology*, 29(8):1225–1251, 2016. 6, 12

[41] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010. 3

[42] Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011. 1, 4

[43] Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In *Artificial intelligence and statistics*, pages 488–495. PMLR, 2009. 1, 4

[44] Yucheng Shi, Kaixiong Zhou, and Ninghao Liu. Engage: Explanation guided data augmentation for graph representation learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 104–121. Springer, 2023. 3

[45] Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling*, 56(10):1936–1949, 2016. 6, 12

[46] Mingchen Sun, Kaixiong Zhou, Xin He, Ying Wang, and Xin Wang. Gppt: Graph pre-training and prompt tuning to generalize graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1717–1727, 2022. 3

[47] Mary M Tai. A mathematical model for the determination of total area under glucose tolerance and other metabolic curves. *Diabetes care*, 17(2):152–154, 1994. 12

[48] Xiao-lan Tian, Si-wei Song, Fang Chen, Xiu-juan Qi, Yi Wang, and Qing-hua Zhang. Machine learning-guided property prediction of energetic materials: Recent advances, challenges, and perspectives. *Energetic Materials Frontiers*, 2022. 1

[49] Paroma Varma, Bryan D He, Payal Bajaj, Nishith Khandwala, Imon Banerjee, Daniel Rubin, and Christopher Ré. Inferring generative model structure with static analysis. *Advances in neural information processing systems*, 30, 2017. 1

[50] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *CoRR*, abs/1710.10903, 2017. 6

[51] Yili Wang, Kaixiong Zhou, Ninghao Liu, Ying Wang, and Xin Wang. Efficient sharpness-aware minimization for molecular graph transformer models. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[52] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding

rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988. 3

[53] 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. 6

[54] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 6, 12

[55] Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, and Jian Tang. Self-supervised graph-level representation learning with local and global structure. In *International Conference on Machine Learning*, pages 11548–11558. PMLR, 2021. 6

[56] Ran Xu, Yue Yu, Hejie Cui, Xuan Kan, Yanqiao Zhu, Joyce C. Ho, Chao Zhang, and Carl Yang. Neighborhood-regularized self-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10611–10619, 2023. 1

[57] Yi Xu, Jiandong Ding, Lu Zhang, and Shuigeng Zhou. Dp-ssl: Towards robust semi-supervised learning with a few labeled samples. *Advances in Neural Information Processing Systems*, 34:15895–15907, 2021. 3

[58] Youjun Xu, Kangjie Lin, Shiwei Wang, Lei Wang, Chenjing Cai, Chen Song, Luhua Lai, and Jianfeng Pei. Deep learning for molecular generation. *Future medicinal chemistry*, 11(6):567–597, 2019. 4

[59] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020. 6

[60] Kaixiong Zhou, Soo-Hyun Choi, Zirui Liu, Ninghao Liu, Fan Yang, Rui Chen, Li Li, and Xia Hu. Adaptive label smoothing to regularize large-scale graph training. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 55–63. SIAM, 2023. 3

[61] Kaixiong Zhou, Zhenyu Zhang, Shengyuan Chen, Tianlong Chen, Xiao Huang, Zhangyang Wang, and Xia Hu. Quangcn: Noise-adaptive training for robust quantum graph convolutional networks. *arXiv preprint arXiv:2211.07379*, 2022. 1