ML-GOOD: Towards Multi-Label Graph Out-Of-Distribution Detection

Tingyi Cai[1,2], **Yunliang Jiang**[2,1,3*], **Ming Li**[4,2], **Changqin Huang**[2], **Yi Wang**[1,2], **Qionghao Huang**[2]

[1]School of Computer Science and Technology, Zhejiang Normal University, China
[2]Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, China
[3]School of Information Engineering, Huzhou University, China
[4]Zhejiang Institute of Optoelectronics, China
tingyicai@zjnu.edu.cn, jyl2022@zjnu.cn, mingli@zjnu.edu.cn, cqhuang@zju.edu.cn, {wangyi, qhhuang}@zjnu.edu.cn

Abstract

The out-of-distribution (OOD) detection on graph-structured data is crucial for deploying graph neural networks securely in open-world scenarios. However, existing methods have overlooked the prevalent scenario of multi-label classification in real-world applications. In this work, we investigate the unexplored issue of OOD detection within multi-label node classification tasks. We propose ML-GOOD, a simple yet sufficient approach that utilizes an energy function to gauge the OOD score for each label. We further develop a strategy for amalgamating multiple label energies, allowing for the comprehensive utilization of label information to tackle the primary challenges encountered in multi-label scenarios. Extensive experimentation conducted on seven diverse sets of real-world multi-label graph datasets, encompassing cross-domain scenarios. The results show that the AUROC of ML-GOOD is improved by 5.26% in intra-domain and 6.54% in cross-domain compared to the previous methods. These empirical validations not only affirm the robustness of our methodology but also illuminate new avenues for further exploration within this burgeoning field of research.

Code — https://github.com/ca1man-2022/ML-GOOD

1 Introduction

Graph neural networks (GNNs) have been widely applied in various fields, ranging from fraud detection (Ma et al. 2023) to medical diagnosis (Xu et al. 2024). However, models deployed in open-world scenarios often encounter out-of-distribution (OOD) input samples from different distributions that were not seen during training (Chen et al. 2022; Shen et al. 2024; Sui et al. 2023). Ideally, a trustworthy GNN model should not only generate accurate predictions for in-distribution (IND) data but also be able to distinguish and reject OOD samples without further prediction.

With the wide application of GNN, the problem of how to recognize graph OOD samples has begun to attract the attention of researchers (Ju et al. 2024; Yu, Liang, and He 2023; Guérin et al. 2023), and various graph OOD detection methods have recently emerged (Li et al. 2022; Guo et al. 2023; Wang et al. 2024b). However, these studies primarily focus on detecting OOD samples in the context of multi-class classification, where each sample is assigned to a single label. This approach is impractical for many real-world applications (Zhao et al. 2023; Huang et al. 2024). For instance, in protein-protein interaction networks, each protein is labeled with multiple functions or associated with various diseases. Similarly, in social networks, identifying anomalous users can involve detecting multiple interest labels linked to each user (Wang, Cui, and Zhu 2016). The transition from single-label to multi-label OOD detection is not as straightforward as it may seem. Multi-label scenarios introduce a unique set of complexities. Figure 1.(a) visually delineates the complexities inherent in the transition from single-label to multi-label OOD detection. In multi-label node classification tasks, such as those found in protein prediction, the initial challenge is to accurately estimate the uncertainty for each node, as shown in Figure 1.(b).

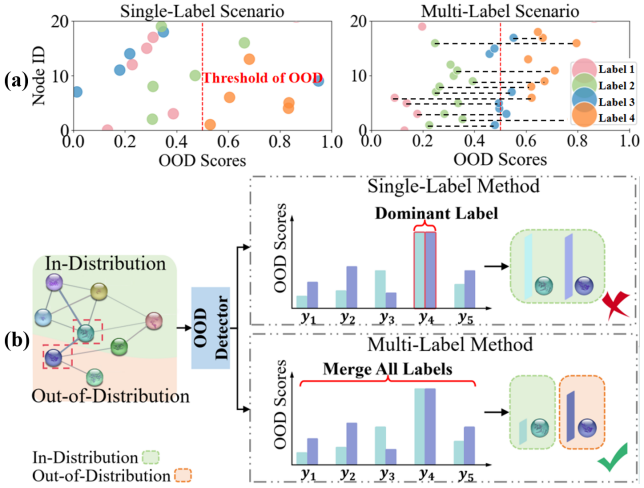

Figure 1: OOD detection for multi-label node classification. The OOD score of each label of a node is computed by a OOD detector on graph data with distributional differences.

If the detection process relies solely on information from a dominant label, it may incorrectly deduce that two nodes with similar OOD scores for that label belong to the same distribution, resulting in substantial estimation bias. Effective estimation of sample uncertainty in a multi-label con-

*Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

text requires simultaneously considering information from multiple labels. In other words, OOD uncertainty prediction must account for a node's relationships with multiple labels to avoid biases stemming from reliance on a single dominant label. In multi-label graph node classification, the OOD detection process must consider the likelihood that a node could be associated with multiple labels. This necessitates that OOD detection methods are capable of not only identifying potential unknown categories but also of assessing the strength of association between the node and any known categories. **However, there remains a significant gap in the development of OOD detection algorithms tailored for multi-label node classification tasks.** To bridge this gap, our work introduces a precise and holistic strategy for identifying OOD samples in multi-label graph node classification. To the best of our knowledge, we are the first to address the issue of multi-label graph OOD detection.

In this paper, we address the crucial issue of OOD uncertainty estimation in multi-label classification settings and propose a simple yet sufficient method for jointly characterizing uncertainty across multiple labels, circumventing the challenging optimization process inherent in training generative models. This method, named Towards **M**ulti-**L**abel **G**raph **O**ut-**O**f-**D**istribution Detection (ML-GOOD), is validated through extensive experiments, demonstrating its superior performance in multi-label graph OOD detection. We summarize our primary contributions:

- **New Issue for Graph OOD.** We are the first to address OOD detection in graph data from a multi-label viewpoint and define a partitioning criterion for the distribution of multi-label datasets.
- **A Simple yet Strong Baseline.** We introduce ML-GOOD, a straightforward and potent technique for measuring uncertainty in multi-label network classifications. Our comprehensive experiments include five single datasets and two cross-domain data groups, ultimately proving its excellent performance.
- **Theoretical Foundation of ML-GOOD.** We offer a clear rationale for effectiveness of ML-GOOD in OOD detection for multi-label graphs. Employing mathematical reasoning, we demonstrate our method's practicality and reliability.

2 Related Work and Preliminaries

In this section, we first present related work on multi-label node classification and graph OOD detection. Subsequently, we introduce the preliminaries of energy-based models.

2.1 Multi-Label Node Classification

In the domain of multi-label node classification (Zhao et al. 2023; Huang et al. 2023; Wang et al. 2024c; Ma, Xu, and Rong 2024), researchers are focused on assigning multiple labels to each instance, which is crucial in fields like text categorization and protein function prediction. Existing approaches (Akujuobi et al. 2019; Shi et al. 2022; Zhou et al. 2021; Song et al. 2021; Wang et al. 2020; You et al. 2020; Xu et al. 2024; Pliakos, Vens, and Tsoumakas 2021) leverage techniques such as reinforcement learning, attention mechanisms, and label embeddings to enhance classification performance. While these methods excel in handling samples from the training distribution, they often struggle with OOD samples, which are critical in safety-critical applications like medical diagnosis. Thus, there is an ongoing exploration of OOD detection methods tailored for multi-label node classification tasks.

2.2 Graph OOD Detection

Graph OOD detection methods have been developed to identify OOD samples effectively. Existing work has developed a variety of GNN OOD detection methods, including techniques based on graph contrastive learning, multi-head attention mechanisms, graph generation models, data-centered frameworks, energy functions, and information bottleneck principles (Liu et al. 2023; Song and Wang 2022; Li et al. 2022; Guo et al. 2023; Wu et al. 2023; Wang et al. 2024b; Stadler et al. 2021; Zhao et al. 2020; Wang et al. 2024a; Bazhenov et al. 2022; Huang, Wang, and Fang 2022; Shen et al. 2024). However, most methods focus on single-label classification, leaving a gap in multi-label OOD research.

2.3 Preliminaries on Energy-Based Model

Energy-based models (EBMs) (Yann et al. 2006) are statistical frameworks that model variable configurations through an energy function $E(\mathbf{x})$. The goal in learning is to align correct configurations with low energy and incorrect ones with high energy. Discriminative models (Liu et al. 2020), like the neural classifier $f(\mathbf{x})$, map inputs to logits, which are then transformed into a probability distribution via the softmax function:

$$p(y|\mathbf{x}) = \frac{e^{f_y(\mathbf{x})/\tau}}{\sum_i^k e^{f_i(\mathbf{x})/\tau}}, \tag{1}$$

where $f_y(\mathbf{x})$ indicates the logit corresponding to the y-th class label and τ is the temperature parameter. This is applicable to single-label classification. EBMs use logits to define a probability density $p(\mathbf{x})$ through the Boltzmann distribution:

$$p(y|\mathbf{x}) = \frac{e^{-E(\mathbf{x},y)/\tau}}{\int_{y'} e^{-E(\mathbf{x},y')/\tau}} = \frac{e^{-E(\mathbf{x},y)/\tau}}{e^{-E(\mathbf{x})/\tau}}, \tag{2}$$

with $E(\mathbf{x}, y) = -f_y(\mathbf{x})$. The free energy function $E(\mathbf{x}; f)$ is derived from the softmax denominator:

$$E(\mathbf{x}; f) = -\tau \cdot \log \sum_i^k e^{f_i(\mathbf{x})/\tau}. \tag{3}$$

3 Problem Formulation

For the task of multi-label classification of graph-structured data, given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A}, \mathbf{X}, \mathbf{Y})$ with $\mathcal{V} = \{v_1, v_2, \cdots, v_n\}$ denoting the set of vertices and $\mathcal{E}$ denoting the set of links/edges between the vertices, then we have $(v_i, v_o) \in \mathcal{E}$. $\mathbf{A} \in \{0, 1\}^{n \times n}$ denotes the adjacency matrix of the graph where $a_{i,o}$ denotes whether there is an edge

connecting nodes v_i and v_o. $\mathbf{X} \in \mathbb{R}^{n \times d}$ represents the feature matrix with n nodes and each node has a feature dimension of d. Furthermore, we denote the label matrix by $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n\}$ and $\mathbf{y}_i = [y_{i1}, y_{i2}, \cdots, y_{ik}]$ denotes the matrix of k labels assigning to node v_i, where $i = 1, 2, \cdots, n$. We can define the objective of the graph out-of-distribution (OOD) detection task within the context of a multi-label setup as:

Definition 1 (Multi-Label Graph OOD Detection). *In scenarios involving real-world graph datasets, we typically encounter two primary categories: those characterized by a single graph and those with multiple graphs. When working with a single-graph dataset, the primary goal of OOD detection is to precisely predict classes for data within the known distribution, while abstaining from making predictions for data identified as OOD. Conversely, within the context of a multi-graph dataset, the overarching objective closely resembles that of a single-graph dataset. Therefore, the objective for multi-label graph OOD detection can be concisely formulated as follows:*

$$\hat{y}_{ij} = \begin{cases} f(\mathbf{x}_i, \mathcal{G})_{[j]}, & \text{if } \mathbf{x}_i \sim D_{in}, \\ \text{reject}, & \text{if } \mathbf{x}_i \sim D_{out}, \end{cases} \tag{4}$$

$$\hat{y}_{ij} = \begin{cases} f(\mathbf{x}_i, \mathcal{G})_{[j]}, & \text{if } (\mathbf{x}_i, \mathcal{G}) \sim D_{in}, \\ \text{reject}, & \text{if } (\mathbf{x}_i, \mathcal{G}) \sim D_{out}, \end{cases} \tag{5}$$

where $\mathbf{x}_i$ is the feature of node v_i, classifier $f(\cdot)$ predicts the j-th label of node v_i and $j = 1, 2, \cdots, k$. To simplify the expression, we uniformly consider the case of multi-graph data below.

Also, when $k = 1$, i.e., in the single-label scenario, the problem can be described as:

$$\hat{\mathbf{y}}_i = \begin{cases} f(\mathbf{x}_i, \mathcal{G}), & \text{if } \mathbf{x}_i \sim D_{in}, \\ \text{reject}, & \text{if } \mathbf{x}_i \sim D_{out}, \end{cases} \tag{6}$$

$$\hat{\mathbf{y}}_i = \begin{cases} f(\mathbf{x}_i, \mathcal{G}), & \text{if } (\mathbf{x}_i, \mathcal{G}) \sim D_{in}, \\ \text{reject}, & \text{if } (\mathbf{x}_i, \mathcal{G}) \sim D_{out}. \end{cases} \tag{7}$$

Therefore, the problem definition we proposed is comprehensive, covering all scenarios of graph OOD detection, ranging from single to multiple labels and from single to multiple graphs.

4 Methodology

In this section, We propose ML-GOOD (**M**ulti-**L**abel **G**raph **O**ut-**O**f-**D**istribution Detection) for identifying OOD instances in multi-label graph classification, where inputs may have multiple labels. As shown in Figure 2, GNNs capture intrinsic graph relationships, while label-specific energy learning generates energy scores per label. Unified and typical energy aggregation methods compute the final sample score for OOD classification, leveraging multi-label information to enhance OOD uncertainty estimation. Furthermore, we provide rigorous mathematical analysis and theoretical support further validate our approach.

4.1 Label-Specific Energy Learning for Graph

To represent graph-structured data with energy, we begin by extracting features from the graph structure data. Subsequently, we map the output logits to energy scores through the specific-label energy function.

Feature Extraction In graph-structured data, nodes represent entities, and edges represent the relationships between these entities. Unlike the pixel grid structure of image data, the complexity of graph data arises from the diversity of nodes and the richness of relationships. In multi-label graph node classification, each node may be associated with multiple labels, reflecting the node's multiple attributes or roles. The connectivity patterns and network topology of graph data reveal additional insights not present in traditional settings with independent and identically distributed (i.i.d.) data. Inspired by (Wu et al. 2023), we leverage Graph Convolutional Networks (GCNs) (Kipf and Welling 2017) to capture the dependencies among nodes. GCNs excel at capturing the nuanced interdependencies among nodes by recursively consolidating and transforming the feature representations of neighboring nodes into enriched node-level representations. Specifically, the feature representation of node v_i at layer l can be described as:

$$\mathbf{h}_{v_i}^{(l)} = \phi(\mathbf{h}_{v_i}^{(l-1)}, \text{AGG}(\{\mathbf{h}_u^{(l-1)}\}_{u \in \mathcal{N}(v_i)})), \tag{8}$$

where $\text{AGG}(\cdot)$ is the aggregation function that pools information from the local neighborhood, $\phi(\cdot)$ is a nonlinear activation function, and $\mathcal{N}(v)$ is the set of 1-hop neighbors of node v. With L layers of graph convolution, the GCNs outputs a vector $\mathbf{h}_{v_i}^{(L)}$ as logits for node v_i, denoted as $h_{\mathbf{x}_i, \mathcal{G}} = \mathbf{h}_{v_i}^{(L)}$. This vector forms the basis for our subsequent computation of energy scores, which is crucial for capturing the unique, multi-dimensional relationships inherent in graph-structured data.

Label-Specific Energy Learning Label-specific energy learning is a foundational concept in our approach that allows us to quantify the uncertainty associated with each label distribution domain. Central to this approach is the specific-label energy function, designed to capture the unique characteristics of graph-structured data. In contrast to conventional single-label classification tasks, multi-label node classification in graph data is inherently more complex. It is structured as a series of binary classification subtasks, each corresponding to a specific label. The goal is to predict the presence or absence of each label for a given node. To achieve this, the final layer of the neural network model typically incorporates a sigmoid layer for multi-label classification, as opposed to the softmax function typically used in single-label classification (Shi et al. 2022). The sigmoid layer transforms the raw model output into probabilities, with each output corresponding to the likelihood of the presence of a specific label:

$$p(y_{ij}|\mathbf{x}_i, \mathcal{G}) = \frac{1}{1 + e^{E(\mathbf{x}_i, \mathcal{G}; j)}} = \frac{e^{h^j_{\mathbf{x}_i, \mathcal{G}}}}{1 + e^{h^j_{\mathbf{x}_i, \mathcal{G}}}}, \tag{9}$$

where y_{ij} is the j-th label of node v_i and $i = 1, 2, \cdots, n$; $j = 1, 2, \cdots, k$. For each binary label y_{ij}, we hypothesize

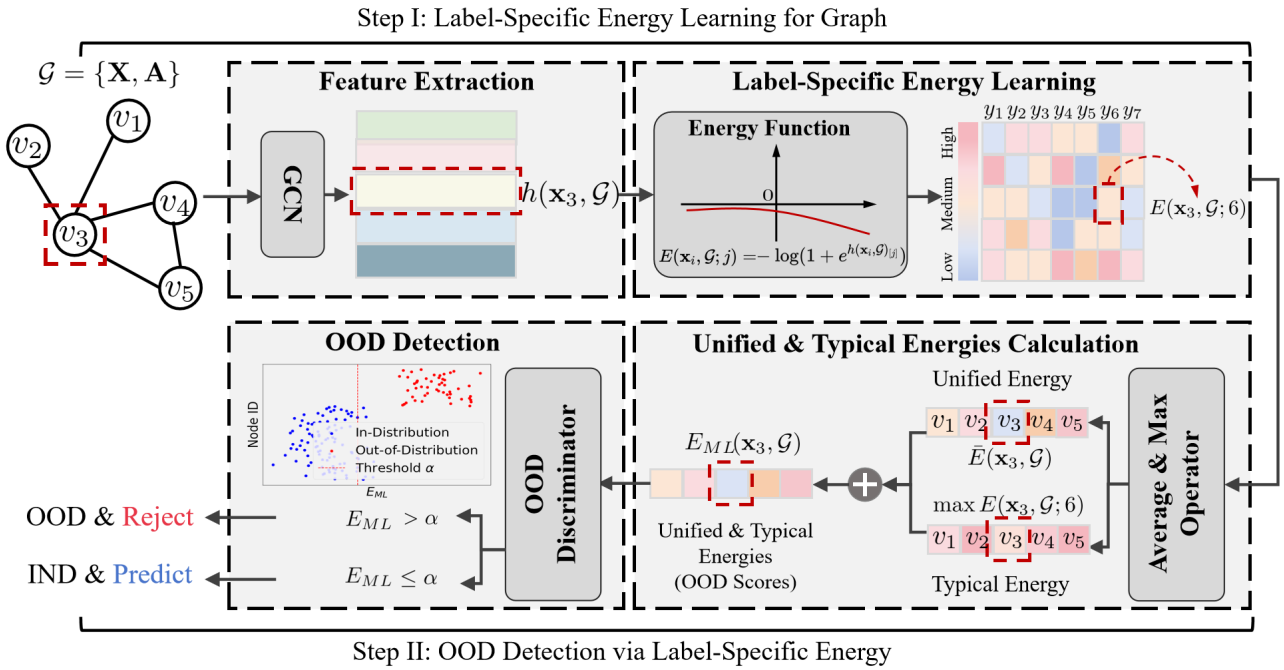

Figure 2: The overall framework of the ML-GOOD method is divided into two steps: Label specific energy learning for graphs (Step I) and OOD detection via label-specific energy learning function (Step II). Step I : Label-specific energy learning for graphs. Firstly, we utilize GCN to capture correlations between graph data, obtaining initial energy scores for each label of the nodes through label-specific energy learning. Step II: OOD detection via label-specific energy function. Initially, unified and typical energy scores are computed, followed by score aggregation to derive the OOD score (i.e., E_{ML}) for nodes. Finally, a threshold α is applied to determine if a node belongs to an OOD sample, consistent with our proposed problem definition.

an energy function $E(\mathbf{x}_i, \mathcal{G}; j) = -h^j_{\mathbf{x}_i, \mathcal{G}}$, where $\mathbf{x}_i$ represents the feature vector of node v_i and $\mathcal{G}$ encapsulates the graph structure. The term $h^j_{\mathbf{x}_i, \mathcal{G}}$ corresponds to the logit for label y_{ij}, derived from the node's feature representation and the graph's topology. By defining label-specific energy function in this manner, we establish a framework to quantify the uncertainty of each label's presence. The free energy function, which incorporates the logit in a logarithmic form, is defined as:

$$E(\mathbf{x}_i, \mathcal{G}; j) = -\log(1 + e^{h^j_{\mathbf{x}_i, \mathcal{G}}}). \tag{10}$$

This formulation, utilizing the logarithmic function, ensures that the energy score $E(\mathbf{x}_i, \mathcal{G}; j)$ is negative, reflecting the inverse relationship between energy and the likelihood of the label's association with the node.

4.2 OOD Detection via Label-Specific Energy

In a subsequent step, we implement a combined approach that integrates unified and typical energy aggregation methods. This strategy allowed us to calculate a final energy score for each sample as an OOD score. In addition, we support our approach through rigorous theoretical analysis.

Unified and Typical Energies Calculation Based on the aforementioned foundation, to capture the dependencies among labels assigned to nodes, we aggregate the energy scores of all labels for each node by computing a unified energy value. This involves calculating the mean of the label-specific energy values across all labels for all nodes. However, directly averaging the multi-label information of nodes may overly smooth the label information, leading to the neglect of high-energy labels while highlighting low-energy ones. To address this issue, we propose incorporating the typical label information (i.e., typical energy) for each node, which entails additionally considering the label with the highest energy score within each node. Specifically, the energy calculation method of ML-GOOD can be expressed as follows:

$$\begin{aligned} E_{ML}(\mathbf{x}_i, \mathcal{G}) &= \bar{E}(\mathbf{x}_i, \mathcal{G}) + \max E(\mathbf{x}_i, \mathcal{G}; j) \\ &= -\frac{1}{k}\sum_{j=1}^{k} \log(1 + e^{h^j_{\mathbf{x}_i, \mathcal{G}}}) \\ &\quad - \log(1 + e^{\min_j h_{\mathbf{x}_i, \mathcal{G}}}). \end{aligned} \tag{11}$$

Finally, the distribution domain of the nodes is discerned using the energy scores obtained from ML-GOOD as the OOD scores of each node. We summarize the complete algorithm for ML-GOOD in Appendix [1].

[1]The appendix is accessible via the provided GitHub link.

4.3 Optimization Objective

The multi-label node classification task is structured as a binary classification task encompassing multiple labels. We adopt the binary cross-entropy loss as the training objective:

$$\mathcal{L}_{bce} = -\frac{1}{N}\sum_{i=1}^{N} y_{ij}\cdot\log(p(y_{ij}))+(1-y_{ij})\cdot\log(1-p(y_{ij})), \tag{12}$$

where $p(y_{ij})$ is the probability that the output belongs to the label y_{ij}, and N is the number of groups of model predictors.

While energy scores are beneficial for pre-trained neural networks, the energy gap between in-distribution and out-of-distribution data may not always be optimal for differentiation. To address this limitation, we partition a subset of the data as OOD training data $\mathcal{D}_{out}^{train}$, which will not be present in the OOD test data, and introduce a regularization technique to create an energy gap. Specifically, the training objectives are as follows:

$$\begin{aligned}\mathcal{L}_{gap} = &\mathbb{E}_{(\mathbf{x}_{in},\mathbf{y})\sim\mathcal{D}_{in}^{train}}(\text{ReLU}(E_{ML}^{in} - m_{in})^2) \\ &+ \mathbb{E}_{(\mathbf{x}_{out},\mathbf{y})\sim\mathcal{D}_{out}^{train}}(\text{ReLU}(m_{out} - E_{ML}^{out})^2),\end{aligned} \tag{13}$$

where m_{in} and m_{out} are the margin hyperparameters, constraining the energy scores of in-distribution and out-of-distribution data to be lower than m_{in} and higher than m_{out}, respectively.

Therefore, the final optimization objective of our model can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_{bce} + \lambda\cdot\mathcal{L}_{gap}, \tag{14}$$

where the parameter λ acts as a regularization factor that adjusts the balance between the contributions of the losses L_{bce} and L_{gap}.

5 Theoretical Analysis

We present the exploration of graph OOD detection through the novel perspective of multi-label classification. A solid theoretical foundation is essential for legitimizing and understanding the methodology we introduce. This section delves into the theoretical framework that supports our proposed method, ML-GOOD, showcasing its capacity to handle the intricacies of multi-label OOD detection through two theorems.

Theorem 1. *Given energy function for i-th node with respect to label j in graph $\mathcal{G}$, and let $h^j_{\mathbf{x}_i,\mathcal{G}}$ denote the corresponding logit. The Multi-Label OOD score for i-th node in graph $\mathcal{G}$ is given by:*

$$E_{ML}(\mathbf{x}_i,\mathcal{G}) \propto \frac{1}{\log p(\mathbf{x}_i|y_{ij}=1)}. \tag{15}$$

For the proofs, please refer to Appendix B. This theorem encapsulates the inverse relationship between the energy of an instance and the density of its label set. We elucidate that ML-GOOD offers a straightforward mechanism for quantifying OOD uncertainty, which is in harmony with the goal of OOD detection. It is explicitly demonstrated that our method awards higher scores to OOD instances, successfully meeting the requisite detection standards.

Based on Theorem 1, we can derive a joint likelihood expression form for E_{ML}. Theorem 2 not only strengthens the theoretical validity of the method, but also highlights its adaptability to the complexity of multi-label OOD detection.

Theorem 2. *For a given node $\mathbf{x}$ and its set of labels $\mathbf{y}_1 = \{y_{11}, y_{12}, \cdots, y_{1k}\}$, the ML-GOOD OOD score $E_{ML}(\mathbf{x},\mathcal{G}) = f(T)$ contains a term T which is the joint likelihood of the conditional probabilities of all labels in the set of labels, i.e.,*

$$T = \prod_{y\in\mathbf{y}_1} p(y|\mathbf{x}). \tag{16}$$

The expression captures joint label estimation, showcasing the efficient use of multi-label information of ML-GOOD. Formal proofs and a discussion on why EBMs suit OOD detection are provided in Appendix B.

6 Experiments

We demonstrate the superior performance of ML-GOOD superior performance on five real-world and two cross-domain datasets, surpassing others on most metrics. Parameter analysis and result visualizations validate our approach, with additional experiments and setups detailed in Appendix A.

6.1 Datasets and Experimental Settings

In this experiment, we employ 6 real-world multi-label datasets including five molecular datasets, `OGB-Proteins`, (Hu et al. 2020), `PPI`, (Zitnik and Leskovec 2017), `HumLoc` and `EukLoc` (Zhao et al. 2023) and one citation network `PCG`.

For the sake of experimental comprehensiveness, we amalgamated two datasets from the similar domain into dataset pairs. Consequently, we conducted two sets of cross-dataset experiments: `PPI+PCG` and `HumLoc+EukLoc`. For datasets `DBLP`, `PCG`, `HumLoc`, and `EukLoc`, which are single-graph datasets lacking obvious domain information, we utilize *feature interpolation*, a method introduced by (Wu et al. 2023) for generating OOD data.

We uniformly use a 2-layer GCN (Kipf and Welling 2017) model as backbone encoder. We use the Adam optimizer (Kingma and Ba 2015) for optimization. The weight decay is 0.01 and learning rate is 0.01. All results are reported on the commonly used metrics (Hendrycks, Mazeika, and Dietterich 2018; Li et al. 2022) AUROC, AUPR, FPR95, and IND ACC. All experimental procedures are conducted on a NVIDIA RTX A6000 GPU device with 48 GB memory.

6.2 Competitors

We compared ML-GOOD with 6 baseline methods, including visual domain methods assuming inputs are independent and identically distributed (i.i.d.): MSP (Hendrycks and Gimpel 2016), ODIN (Liang, Li, and Srikant 2018), Mahalanobis (Lee et al. 2018), OE (Hendrycks, Mazeika, and Dietterich 2018), and Energy Fine-Tune (Liu et al. 2020). We also compared the GNNSafe++ (Wu et al. 2023) approach, a specialized OOD detection method for graph data

Methods	OGB-Proteins				PPI			
	AUROC	AUPR	FPR95	IND ACC	AUROC	AUPR	FPR95	IND ACC
MSP	49.91	44.43	94.95	41.13	49.62	46.99	95.20	45.69
ODIN	50.60	44.55	95.05	41.13	50.10	47.54	94.90	45.72
Mahalanobis	49.85	44.38	94.97	52.78	46.50	45.79	95.16	**51.48**
Energy FT	50.15	44.60	**94.79**	39.41	48.41	46.65	94.81	48.77
OE	49.89	44.42	95.05	42.26	50.38	47.55	94.90	41.11
GNNSafe++	50.36	44.73	94.80	39.23	43.77	44.09	96.88	48.66
ML-GOOD	**50.71**	**44.86**	94.97	**60.21**	**54.50**	**49.58**	**91.81**	47.83

Table 1: OOD detection results in terms of AUROC (↑) / AUPR (↑) / FPR95 (↓) on datasets OGB-Proteins and PPI. The best results are highlighted with **bold**, while suboptimal results are underlined.

Methods	DBLP				PCG				HumLoc			
	AUROC	AUPR	FPR95	IND ACC	AUROC	AUPR	FPR95	IND ACC	AUROC	AUPR	FPR95	IND ACC
MSP	46.64	52.18	99.63	75.40	54.74	50.08	93.88	49.96	58.66	53.39	93.40	62.96
ODIN	53.36	59.54	99.53	75.40	45.26	42.01	97.34	49.96	41.34	39.78	98.23	62.96
Mahalanobis	**83.06**	80.29	**64.46**	89.92	70.67	61.32	73.00	49.23	63.42	54.22	**81.52**	59.75
Energy FT	78.16	76.27	76.83	90.68	54.48	50.94	95.30	45.86	68.94	59.18	89.47	53.91
OE	65.45	67.05	95.44	73.63	71.50	68.45	83.20	**52.97**	69.28	63.15	86.06	**64.40**
GNNSafe++	81.03	83.22	93.04	90.86	90.05	91.46	89.98	46.38	75.10	76.76	95.91	55.95
ML-GOOD	81.84	**86.03**	93.65	**90.89**	**95.31**	**93.52**	**20.94**	48.30	**82.07**	**79.25**	94.40	60.96

Table 2: OOD detection results in terms of AUROC (↑) / AUPR (↑) / FPR95 (↓) on datasets DBLP, PCG and HumLoc with *feature interpolation*. The best results are highlighted with **bold**, while suboptimal results are underlined.

Methods	$\mathcal{D}_{in}$: HumLoc / $\mathcal{D}_{out}$: EukLoc			
	AUROC	AUPR	FPR95	IND ACC
MSP	66.20	33.99	**81.32**	58.24
ODIN	42.42	20.12	94.27	58.24
Energy FT	40.51	19.59	97.70	54.65
OE	43.64	24.61	98.47	53.57
GNNSafe++	61.32	27.88	93.84	55.70
ML-GOOD	**72.65**	**41.84**	90.24	58.27

Table 3: Cross-dataset OOD detection results in terms of AUROC (↑) / AUPR (↑) / FPR95 (↓). The best results are highlighted with **bold**, while suboptimal results are underlined.

that does not assume inputs to be independent and identically distributed (non-i.i.d.).

6.3 Results and Discussion

Table 1 presents the performance evaluation of our method on OGB-Proteins and PPI datasets. Leveraging the structural features of graph data under multi-label conditions, our approach demonstrates superior performance over competing methods across a range of key metrics. Particularly noteworthy is our method's ability to maintain competitive OOD detection performance while preserving excellent in-distribution classification accuracy on the OGB-Proteins dataset. It may be attributed to the fine-grained energy uncertainty estimation method, which proves to be particularly effective for this dataset. Nonetheless, we acknowledge limitations encountered when conducting experiments on large-scale multi-label datasets such as OGB-Proteins, primarily due to memory constraints. To mitigate this, we conducted all experiments by randomly sampling 20% of this dataset, ensuring fairness and representativeness in comparisons.

Table 2 presents the experimental outcomes concerning the three datasets DBLP, PCG, and HumLoc. While retaining comparable IND classification accuracy performance (Fang et al. 2022, 2024) to other methods, our method exhibits excellent performance on a variety of metrics of OOD detection, particularly notable was the significantly low FPR95 metric. The low FPR95 value indicates that our method effectively reduces false positive rates while maintaining high classification accuracy. Furthermore, we observe that the existing literature on OOD detection in graph data has not yet established a consensus regarding the specification of the dataset to be used: whether a single dataset or multiple datasets should be employed. Considering the completeness of the comparison scenarios, we also conducted experiments across datasets. Table 3 presents the OOD detection results for the cross-dataset scenario. The results show that our approach still maintains an advantage in terms of OOD detection performance when crossing domains. Moreover, tailored methods designed for graph-structured data prove to be more effective compared to those assuming input independence and uniformity. In summary, the analysis of experimental results underscores the robustness and effectiveness of our approach in discerning between IND and OOD samples.

Methods	OGB-Proteins		PPI		DBLP		PCG		HumLoc	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
ML-GOOD(w/o LS)	49.91	44.20	50.98	47.75	81.45	84.22	66.31	66.90	63.98	67.60
ML-GOOD(w/ LS)	**50.71**	**44.86**	**54.50**	**49.58**	**81.84**	**86.03**	**95.31**	**93.52**	**82.07**	**79.25**

Table 4: Ablation study results in terms of AUROC(↑) / AUPR (↑). The "LS" stands for label-specific energy score calculation. The best performing method in each experiment is in **bold**.

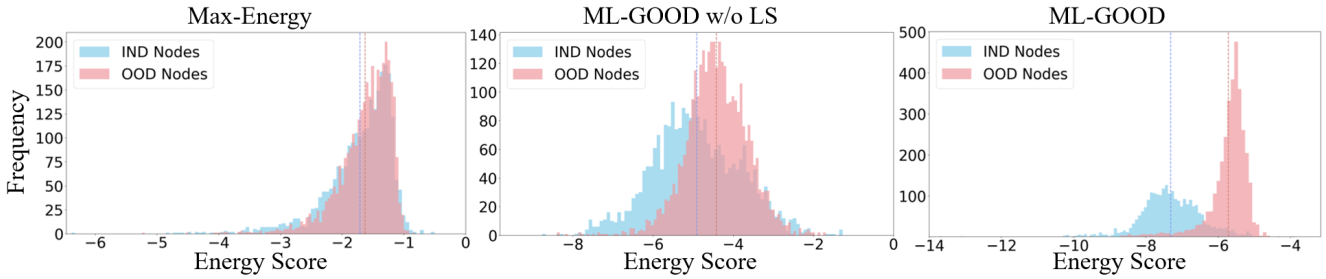

Figure 3: The energy distributions on PCG. The out-of-distribution nodes (red) have higher scores.

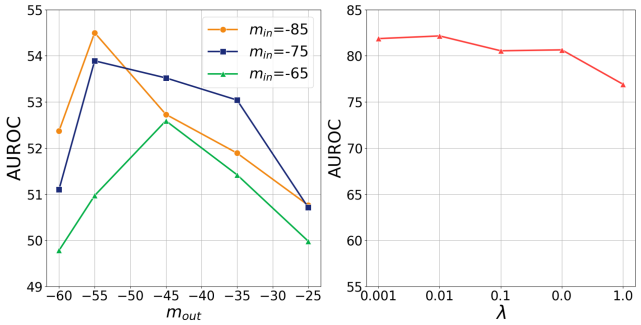

Figure 4: Hyperparameters analysis. Left: impact of the $m_{out} \in \{-60, -55, -45, -35, -25\}$ and $m_{in} \in \{-85, -75, -65\}$ margin hyperparameters on PPI; Right: impact of the weight hyperparameter λ on DBLP.

6.4 Ablation Study

To validate the proposed energy function computation for the multi-label scenario, we performed ablation experiments, with the findings presented in Table 4. The "LS" stands for label-specific energy score calculation. The results indicate that taking into account multi-label information comprehensively can significantly enhance the effectiveness of graph multi-label OOD detection.

6.5 Parameter Sensitivity Analysis

We examine the impact of the margin hyperparameters m_{out} and m_{in}, along with the regularization weight λ. The outcomes are shown in Figure 4. In establishing the margin hyperparameters, we initially derive their values from data exhibiting varied distributions and subsequently identify the optimal parameters by assessing the performance trend and expanding the energy gap from these initial values. The results highlight the significant role margin hyperparameters in the effectiveness of ML-GOOD, as our detection strategy relies on identifying variations in energy scores across different data distributions. Notably, prudent adjustment of regularization weights λ is essential, as excessive augmentation of regularization intensity may engender performance deterioration.

6.6 Visualization

To visually illustrate the impact of our methods, we present the distribution of energy scores for IND and OOD inputs, as shown in Figure 3. Our findings demonstrate that the label-specific calculation effectively discriminates the energy differential between IND and OOD data, thereby indicating the intuitive capability of ML-GOOD in detecting OOD nodes. Furthermore, this observation substantiates our initial hypothesis regarding the existence of nodes concurrently associated with multiple labels, and underscores the importance of considering the energy scores corresponding to each label. Such consideration offers a more nuanced depiction of the underlying complexity in relationships, an aspect potentially overlooked when solely focusing on the maximum energy score. Thus, our results affirm the effectiveness of the proposed label-specific energy score calculation and regularization approach.

7 Conclusion

In this paper, we tackle the emerging issue of multi-label graph out-of-distribution (GOOD) detection. We propose a label-specific energy function that effectively integrates multi-label information to compute GOOD energy scores, supported by theoretical analysis. Through extensive experimentation and discussion, we verify the efficacy of our approach, i.e., ML-GOOD. We aim to inspire further research into OOD detection for multi-label graph classification.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. U22A20102, 62337001, 62172370, U21A20473), the Jinhua Science and Technology Plan (No. 2023-3-003a), and the National Natural Science Foundation of China (Grant No. 62207028).

References

Akujuobi, U.; Yufei, H.; Zhang, Q.; and Zhang, X. 2019. Collaborative Graph Walk for Semi-Supervised Multi-label Node Classification. In *ICDM*, 1–10.

Bazhenov, G.; Ivanov, S.; Panov, M.; Zaytsev, A.; and Burnaev, E. 2022. Towards OOD detection in Graph Classification From Uncertainty Estimation Perspective. In *ICML*.

Chen, Y.; Zhang, Y.; Bian, Y.; Yang, H.; Kaili, M.; Xie, B.; Liu, T.; Han, B.; and Cheng, J. 2022. Learning Causally Invariant Representations for Out-of-Distribution Generalization on Graphs. In *NeurIPS*, 22131–22148.

Fang, Z.; Li, Y.; Liu, F.; Han, B.; and Lu, J. 2024. On the Learnability of Out-of-Distribution Detection. *Journal of Machine Learning Research*, 25: 1–83.

Fang, Z.; Li, Y.; Lu, J.; Dong, J.; Han, B.; and Liu, F. 2022. Is Out-of-Distribution Detection Learnable? 37199–37213.

Guérin, J.; Delmas, K.; Ferreira, R.; and Guiochet, J. 2023. Out-of-Distribution Detection Is Not All You Need. In *AAAI*, 14829–14837.

Guo, Y.; Yang, C.; Chen, Y.; Liu, J.; Shi, C.; and Du, J. 2023. A Data-Centric Framework to Endow Graph Neural Networks With Out-of-Distribution Detection Ability. In *KDD*, 638–648.

Hendrycks, D.; and Gimpel, K. 2016. A Baseline for Detecting Misclassified and Out-of-distribution Examples in Neural Networks. In *ICLR*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018. Deep Anomaly Detection With Outlier Exposure. In *ICLR*.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. In *NeurIPS*, 22118–22133.

Huang, C.; Wang, Y.; Jiang, Y.; Li, M.; Huang, X.; Wang, S.; Pan, S.; and Zhou, C. 2024. Flow2GNN: Flexible Two-Way Flow Message Passing for Enhancing GNNs Beyond Homophily. *IEEE Transactions on Cybernetics*, 54(11): 6607–6618.

Huang, M.; Zhao, Y.; Wang, Y.; Wahab, F.; Sun, Y.; and Chen, C. 2023. Multi-Graph Multi-Label Learning With Novel and Missing Labels. *Knowledge-Based Systems*, 276: 110753.

Huang, T.; Wang, D.; and Fang, Y. 2022. End-to-End Open-Set Semi-Supervised Node Classification With Out-of-Distribution Detection. In *IJCAI*, 23–29.

Ju, W.; Yi, S.; Wang, Y.; Xiao, Z.; Mao, Z.; Li, H.; Gu, Y.; Qin, Y.; Yin, N.; Wang, S.; et al. 2024. A Survey of Graph Neural Networks in Real world: Imbalance, Noise, Privacy and OOD Challenges. *arXiv preprint arXiv:2403.04468*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *NeurIPS*, 7167–7177.

Li, Z.; Wu, Q.; Nie, F.; and Yan, J. 2022. GraphDE: A Generative Framework for Debiased Learning and Out-of-Distribution Detection on Graphs. In *NeurIPS*, 30277–30290.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks. In *ICLR*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-Based Out-of-Distribution Detection. In *NeurIPS*, 21464–21475.

Liu, Y.; Ding, K.; Liu, H.; and Pan, S. 2023. GOOD-D: On Unsupervised Graph Out-of-Distribution Detection. In *WSDM*, 339–347.

Ma, J.; Li, F.; Zhang, R.; Xu, Z.; Cheng, D.; Ouyang, Y.; Zhao, R.; Zheng, J.; Zheng, Y.; and Jiang, C. 2023. Fighting Against Organized Fraudsters Using Risk Diffusion-Based Parallel Graph Neural Network. In *IJCAI*, 6138–6146.

Ma, J.; Xu, F.; and Rong, X. 2024. Discriminative Multi-Label Feature Selection With Adaptive Graph Diffusion. *Pattern Recognition*, 148: 110154.

Pliakos, K.; Vens, C.; and Tsoumakas, G. 2021. Predicting Drug-Target Interactions With Multi-Label Classification and Label Partitioning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18: 1596–1607.

Shen, X.; Wang, Y.; Zhou, K.; Pan, S.; and Wang, X. 2024. Optimizing OOD Detection in Molecular Graphs: A Novel Approach With Diffusion Models. In *KDD*.

Shi, M.; Tang, Y.; Zhu, X.; and Liu, J. 2022. Multi-Label Graph Convolutional Network Representation Learning. *IEEE Transactions on Big Data*, 8: 1169–1181.

Song, Y.; and Wang, D. 2022. Learning on Graphs With Out-of-Distribution Nodes. In *KDD*, 1635–1645.

Song, Z.; Meng, Z.; Zhang, Y.; and King, I. 2021. Semi-Supervised Multi-Label Learning for Graph-Structured Data. In *CIKM*, 1723–1733.

Stadler, M.; Charpentier, B.; Geisler, S.; Zügner, D.; and Günnemann, S. 2021. Graph Posterior Network: Bayesian Predictive Uncertainty for Node Classification. In *NeurIPS*, 18033–18048.

Sui, Y.; Wu, Q.; Wu, J.; Cui, Q.; Li, L.; Zhou, J.; Wang, X.; and He, X. 2023. Unleashing the Power of Graph Data Augmentation on Covariate Distribution Shift. In *NeurIPS*, 18109–18131.

Wang, D.; Cui, P.; and Zhu, W. 2016. Structural Deep Network Embedding. In *KDD*, 1225–1234. ISBN 9781450342322.

Wang, F.; Liu, Y.; Liu, K.; Wang, Y.; Medya, S.; and Yu, P. S. 2024a. Uncertainty in Graph Neural Networks: A Survey. *arXiv preprint arXiv:2403.07185*.

Wang, L.; He, D.; Zhang, H.; Liu, Y.; Wang, W.; Pan, S.; Jin, D.; and Chua, T.-S. 2024b. GOODAT: Towards Test-time Graph Out-of-Distribution Detection. In *AAAI*, 15537–15545.

Wang, Y.; He, D.; Li, F.; Long, X.; Zhou, Z.; Ma, J.; and Wen, S. 2020. Multi-Label Classification With Label Graph Superimposing. In *AAAI*, 12265–12272.

Wang, Y.; Zhao, Y.; Wang, Z.; Zhang, C.; and Wang, X. 2024c. Robust Multi-Graph Multi-Label Learning With Dual-Granularity Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10): 6509–6524.

Wu, Q.; Chen, Y.; Yang, C.; and Yan, J. 2023. Energy-based Out-of-Distribution Detection for Graph Neural Networks. In *ICLR*.

Xu, J.; Yuan, C.; Ma, X.; Shang, H.; Shi, X.; and Zhu, X. 2024. Interpretable Medical Deep Framework by Logits-Constraint Attention Guiding Graph-Based Multi-Scale Fusion for Alzheimer's Disease Analysis. *Pattern Recognition*, 152: 110450.

Yann, L.; Sumit, C.; Raia, H.; Marc'Aurelio, R.; and Fu-Jie, H. 2006. *A Tutorial on Energy-Based Learning*. MIT Press.

You, R.; Guo, Z.; Cui, L.; Long, X.; Bao, Y.; and Wen, S. 2020. Cross-Modality Attention With Semantic Graph Embedding for Multi-Label Classification. In *AAAI*, 12709–12716.

Yu, J.; Liang, J.; and He, R. 2023. Mind the Label Shift of Augmentation-Based Graph OOD Generalization. In *CVPR*, 11620–11630.

Zhao, T.; Dong, T. N.; Hanjalic, A.; and Khosla, M. 2023. Multi-Label Node Classification On Graph-Structured Data. *Transactions on Machine Learning Research*. URL: https://openreview.net/forum?id=EZhkV2BjDP.

Zhao, X.; Chen, F.; Hu, S.; and Cho, J.-H. 2020. Uncertainty Aware Semi-Supervised Learning on Graph Data. In *NeurIPS*, 12827–12836.

Zhou, C.; Chen, H.; Zhang, J.; Li, Q.; Hu, D.; and Sheng, V. S. 2021. Multi-Label Graph Node Classification With Label Attentive Neighborhood Convolution. *Expert Systems with Applications*, 180: 115063.

Zitnik, M.; and Leskovec, J. 2017. Predicting Multicellular Function Through Multi-Layer Tissue Networks. *Bioinformatics*, (14): i190–i198.