Appendix A

Evaluation Metrics

We assess the following commonly utilized metrics for OOD detection (Hendrycks, Mazeika, and Dietterich 2018; Li et al. 2022): area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPR), and the false positive rate at 95% true positive rate (FPR95). The in-distribution performance is measured by the Accuracy on testing nodes, i.e., IND ACC.

Additional Experiments

Tbale 1 presents the OOD detection results for the cross-dataset scenario. The results for the other cross-dataset case are reported as follows: Table 1 presents the results with `PPI` as the in-distribution dataset and `PCG` as the out-of-distribution dataset.

Table 1: Cross-dataset OOD detection results in terms of AUROC (↑) / AUPR (↑) / FPR95 (↓). The best results are highlighted with **bold**, while suboptimal results are underlined.

Methods	$\mathcal{D}_{in}$: PPI / $\mathcal{D}_{out}$: PCG			
	AUROC	AUPR	FPR95	IND ACC
MSP	47.31	28.29	92.70	44.75
ODIN	96.13	79.54	6.43	43.79
Energy FT	<u>97.92</u>	92.73	<u>5.57</u>	**45.15**
OE	73.29	44.67	53.94	44.45
GNNSafe++	96.04	<u>93.76</u>	31.80	<u>45.02</u>
ML-GOOD	**99.90**	**99.82**	**0.03**	44.17

Notation Table

Table 2 summarizes the notations and definitions throughout this paper for clarity.

Hyperparameters Settings

Table 3 reports more hyperparameter settings on the datasets.

Datasets

In this paper, we employ 6 real-world multi-label datasets including `OGB-Proteins` (Hu et al. 2020), `PPI` (Zitnik and Leskovec 2017), `DBLP`, `PCG`, `HumLoc` and `EukLoc` (Zhao et al. 2023). Table 4 presents basic information about the multi-label datasets we used.

- The `OGB-Proteins` dataset represents a protein-protein association network, wherein the objective is to forecast the presence of protein functions in a multi-label binary classification framework, encompassing a total of 112 labels for prediction. In-distribution and out-of-distribution data by species category.
- The `PPI` dataset comprises 20 graphs, each representing a distinct human tissue, with gene ontology annotations serving as labels (a total of 121 labels). We classify the first 8 graphs as in-distribution data and the remaining 12 graphs as out-of-distribution data.
- The `DBLP` dataset is an integrated database system for computer science literature in English, focusing on research outcomes. It features authors at the heart of the system, with 4 labels representing their respective fields.

Table 2: Notation table.

Notation	Description	Notation	Description
$\mathcal{G}$	a graph	$\mathbf{y}_v$	label matrix of vertex
$\mathcal{V}$	vertex set	n	number of nodes
$\mathcal{E}$	links/edges set	d	feature dimension
$\mathbf{A}$	adjacency matrix	l	network layer
v	a vertex	D_{in}	in-distribution data
$\mathbf{X}$	feature matrix	D_{out}	out-of-distribution data
$\mathbf{Y}$	label matrix	N	number of model predictors

Table 3: Margin hyperparameter settings on different datasets.

Datasets	m_{out}	m_{in}
OGB-Proteins	{-45, -35, -25, -15}	{-75, -65, -55}
PCG	{-8, -5, -2, 1}	{-12, -9, -6}
DBLP	{-5, -2, 1, 4}	{-9, -5, -1}
HumLoc	{-12, -6, -3, 0}	{-20, -15, -5}

- The PCG dataset is a specialized collection crafted for the purpose of protein phenotype prediction. It comprises 3,233 nodes and 37,351 edges, and includes a comprehensive set of 15 distinct labels.
- The HumLoc dataset comprises 3,106 nodes and 18,496 edges, and its primary objective is to predict the subcellular locations of human proteins. Each protein can have one to several labels in 14 possible locations.
- The EukLoc dataset features 7,766 proteins as nodes and 13,818 connections as edges for predicting the subcellular location of eukaryotic proteins. Each node is associated with up to 22 different labels.

Table 4: Statistic of the datasets.

Datasets	# Nodes	# Edges	# Graphs	# Labels
OGB-Proteins	132,534	21,446,852	1	112
PPI	44,906	44,906	20	121
DBLP	28,706	68,335	1	4
PCG	3,233	74,702	1	15
HumLoc	3,106	18,496	1	14
EukLoc	7,766	13,818	1	22

Segmentation of the Distribution Domains We have observed that existing literature OOD detection in graph data has yet to reach a consensus on the specification of datasets to be employed: whether a single dataset or multiple datasets. 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 a single dataset, we randomly partition the data into in-distribution training and out-of-distribution test sets based on a predetermined ratio. 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. This involved employing random interpolation to generate node features for the OOD data while retaining the original graphs as the in-distribution data.

Appendix B

Algorithm

Discussion on ML-GOOD

To provide additional insights into the rationale and significance of our proposed ML-GOOD, we analyze the applicability of EBMs in both traditional OOD detection and multi-label graph OOD detection. We address two fundamental questions: 1. *Why are EBMs well-suited for OOD detection?* 2. *Why is ML-GOOD well-suited for multi-label graph OOD detection?*

Why are EBMs well-suited for OOD detection? In the filed of OOD detection, the primary objective is to distinguish between familiar, in-distribution data and unfamiliar, out-of-distribution (OOD) data, which may contain novel or anomalous samples. Ensuring precise quantification of uncertainty becomes paramount for robust OOD detection, particularly when dealing with unseen data instances. Energy-based models (EBMs) (Du and Mordatch 2019; Zhai et al. 2016; Grathwohl et al. 2019; Xie, Zheng, and Li 2021; Nijkamp et al. 2020), deeply rooted in statistical physics, provide a framework for characterizing the likelihood of different states based on their energy levels. The fundamental concept of the Boltzmann distribution directly links energy to probability, favoring lower energy states as more probable occurrences. This intrinsic property aligns with the essence of OOD detection, as OOD samples are expected to exhibit deviations in their energy levels compared to in-distribution data. By leveraging the Boltzmann distribution and the concept of maximum entropy, EBMs present a principled approach to model the uncertainty associated with data points. Computing the energy of a data point using the energy function allows quantification of its compatibility with the learned distribution, with anomalously high energy scores serving as potential indicators of OOD samples. Furthermore, EBMs offer a systematic methodology to calibrate uncertainty estimates. The continuous measure of uncertainty provided by the energy function effectively mirrors the model's confidence in its predictions. Regions characterized by high-energy states are indicative of high uncertainty, suggesting the presence of novel or unfamiliar data. This inherent

Algorithm 1: ML-GOOD: Towards Multi-Label Graph Out-Of-Distribution Detection

Input: A multi-label graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A}, \mathbf{X}, \mathbf{Y})$, threshold for OOD uncertainty α
Output: The results of OOD detetion $\hat{y}$

for $i = 1$ to n **do**
  # Extract feature representation:
  $\mathbf{h}_v^{(l)} = \phi(\mathbf{h}_v^{(l-1)}, \text{AGG}(\{\mathbf{h}_u^{(l-1)}\}_{u \in \mathcal{N}(v)}))$;
  $h_{\mathbf{x}_i, \mathcal{G}} = \mathbf{h}_{v_i}^{(L)}$
  # Compute label-specific energy:
  $E(\mathbf{x}_i, \mathcal{G}; j) = -\log(1 + e^{h^j_{\mathbf{x}_i, \mathcal{G}}})$
  # Compute ML energy:
  $E_{ML}(\mathbf{x}_i, \mathcal{G}) = \bar{E}(\mathbf{x}_i, \mathcal{G}) + \max E(\mathbf{x}_i, \mathcal{G}; j)$
  if $E_{ML}(\mathbf{x}_i, \mathcal{G}) > \alpha$ **then**
    $\hat{y}_i$ = reject
  else
    $\hat{y}_i$ = Sigmoid$(h^j_{\mathbf{x}_i, \mathcal{G}})$
  end if
  return $\hat{y}_i$
end for

capability to quantify uncertainty positions EBMs as well-suited candidates for OOD detection tasks, where robust uncertainty estimates are essential for decision-making.

Why is ML-GOOD well-suited for multi-label graph OOD detection? Motivated by the theoretical insights of (Wang et al. 2021), we undertake theoretical analysis for ML-GOOD. We denote the conditional likelihood $p(\mathbf{x}_i | y_{ij} = 1)$ according to EBM as:

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

and $Z_{y_{ij}} = \int_{\mathbf{x}_i | y_{ij}} e^{-E(\mathbf{x}_i, \mathcal{G}; j)}$ is the normalized densities with respect to $\mathbf{x}_i$. We subsequently reformulate the ML-GOOD expression into a new format with joint likelihood perspective:

$$\begin{aligned}
E_{ML}(\mathbf{x}_i, \mathcal{G}) &= \bar{E}(\mathbf{x}_i, \mathcal{G}) + \max E(\mathbf{x}_i, \mathcal{G}; j) \\
&= \frac{1}{k} \sum_{j=1}^{k} E(\mathbf{x}_i, \mathcal{G}; j) + \max E(\mathbf{x}_i, \mathcal{G}; j) \\
&= -\frac{1}{k} \sum_{j=1}^{k} \log \left(p(\mathbf{x}_i | y_{ij} = 1) \cdot Z_{y_{ij}} \right) + \max_j \left(-\log \left(p(\mathbf{x}_i | y_{ij} = 1) \cdot Z_{y_{ij}} \right) \right) \\
&= -\frac{1}{k} \sum_{j=1}^{k} \log p(\mathbf{x}_i | y_{ij} = 1) \underbrace{-\frac{1}{k} \sum_{j=1}^{k} \log Z_{y_{ij}}}_{C_1} + \max_j \left(-\log p(\mathbf{x}_i | y_{ij} = 1) - \log Z_{y_{ij}} \right) \\
&= \underbrace{-\frac{1}{k} \sum_{j=1}^{k} \log p(\mathbf{x}_i | y_{ij} = 1)}_{T_1} + \underbrace{\max_j \left(-\log p(\mathbf{x}_i | y_{ij} = 1) - \log Z_{y_{ij}} \right)}_{T_2} + C_1,
\end{aligned} \tag{2}$$

where C_1 denotes constant term. Afterwards, by applying the Bayesian theorem transformation for $T_1 = -\frac{1}{k}\sum_{j=1}^{k}\log p(\mathbf{x}_i|y_{ij}=1)$, we obtain:

$$\begin{aligned}
T_1 &= -\frac{1}{k}\sum_{j=1}^{k}\log\frac{p\left(y_{ij}=1|\mathbf{x}_i\right)\cdot p\left(\mathbf{x}_i\right)}{p\left(y_{ij}=1\right)} \\
&= -\frac{1}{k}\sum_{j=1}^{k}\left(\log p(y_{ij}=1|\mathbf{x}_i)+\log p(\mathbf{x}_i)-\log p(y_{ij}=1)\right) \\
&= -\frac{1}{k}\log\prod_{j=1}^{k}p(y_{ij}=1|\mathbf{x}_i)-\log p(\mathbf{x}_i)+\frac{1}{k}\log\prod_{j=1}^{k}p(y_{ij}=1) \\
&= -\frac{1}{k}\log p\left(\left(y_{i1}=1,y_{i2}=1,\cdots,y_{ik}=1\right)|\mathbf{x}_i\right)-\log p(\mathbf{x}_i)+\frac{1}{k}\log\prod_{j=1}^{k}p(y_{ij}=1) \\
&= -\frac{1}{k}\log\left(\frac{p\left(\mathbf{x}_i|y_{i1}=1,y_{i2}=1,\cdots,y_{ik}=1\right)}{p\left(\mathbf{x}_i\right)}\cdot\prod_{j=1}^{k}p(y_{ij}=1)\right)-\log p(\mathbf{x}_i)+\frac{1}{k}\log\prod_{j=1}^{k}p(y_{ij}=1)
\end{aligned} \tag{3}$$

$$\begin{aligned}
&= -\frac{1}{k}\log p\left(\mathbf{x}_i|y_{i1}=1,y_{i2}=1,\cdots,y_{ik}=1\right)-\frac{1}{k}\log\prod_{j=1}^{k}p(y_{ij}=1)+\frac{1}{k}\log p(\mathbf{x}_i)-\log p(\mathbf{x}_i)+\frac{1}{k}\log\prod_{j=1}^{k}p(y_{ij}=1) \\
&= \underbrace{-\frac{1}{k}\log p\left(\mathbf{x}_i|y_{i1}=1,y_{i2}=1,\cdots,y_{ik}=1\right)}_{\uparrow \text{ for OOD}}\underbrace{-\left(1-\frac{1}{k}\right)\log p\left(\mathbf{x}_i\right)}_{\uparrow \text{ for OOD}}.
\end{aligned} \tag{4}$$

For the term $T_2 = \max_j\left(-\log\left(p\left(\mathbf{x}_i|y_{ij}=1\right)\cdot Z_{y_{ij}}\right)\right)$, we have:

$$T_2 = -\log\min\left(p(\mathbf{x}_i|y_{ij}=1)\cdot Z_{y_{ij}}\right). \tag{5}$$

Referring to label-specific energy function and Eq.(1), we can derive:

$$\begin{aligned}
Z_{y_{ij}} &= \int_{\mathbf{x}_i|y_{ij}} e^{-E(\mathbf{x}_i,\mathcal{G};j)} \\
&= \int_{\mathbf{x}_i|y_{ij}} e^{\log(1+e^{h^j_{\mathbf{x}_i,\mathcal{G}}})} > 0.
\end{aligned} \tag{6}$$

Since $1+e^{h^j_{\mathbf{x}_i,\mathcal{G}}}$ is monotonically increasing, its logarithm and exponent are also monotonically increasing functions. Thus, $Z_{y_{ij}}$ is a positive number and as $h^j_{\mathbf{x}_i,\mathcal{G}}$ increases. Since $Z_{y_{ij}}$ is monotonically increasing, it follows that $p(\mathbf{x}_i|y_{ij}=1)\cdot Z_{y_{ij}}$ is also monotonically increasing. When $\min\left(p(\mathbf{x}_i|y_{ij}=1)\cdot Z_{y_{ij}}\right)$ is monotonically increasing, its negative logarithm will be monotonically decreasing. That is, T_2 is monotonically decreasing. This component aligns with the requirement that energy scores correspond to greater uncertainty in OOD nodes.

Ultimately, we arrive at the following outcomes:

$$E_{ML}(\mathbf{x}_i,\mathcal{G}) = \underbrace{T_1}_{\uparrow \text{ for OOD}} + \underbrace{T_2}_{\uparrow \text{ for OOD}} + C_1, \tag{7}$$

which shows that the computational approach employed by ML-GOOD not only conforms to the principle that lower probability densities (i.e., out-of-distribution) correspond to higher scores, which is in line with the requirement for OOD detection, but also, notably, the first term incorporates joint estimation across labels. This suggests that our method effectively captures multi-label information in a straightforward manner, bypassing the complexities associated with optimization processes.

References

Du, Y.; and Mordatch, I. 2019. Implicit generation and modeling with energy based models. In *NeurIPS*, 3608–3618.

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2019. Your classifier is secretly an energy based model and you should treat it like one. 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.

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.

Nijkamp, E.; Hill, M.; Han, T.; Zhu, S.-C.; and Wu, Y. N. 2020. On the anatomy of MCMC-based maximum likelihood learning of energy-based models. In *AAAI*, 5272–5280.

Wang, H.; Liu, W.; Bocchieri, A.; and Li, Y. 2021. Can multi-label classification networks know what they don't know? In *NeurIPS*, 29074–29087.

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

Xie, J.; Zheng, Z.; and Li, P. 2021. Learning energy-based model with variational auto-encoder as amortized sampler. In *AAAI*, 10441–10451.

Zhai, S.; Cheng, Y.; Lu, W.; and Zhang, Z. 2016. Deep structured energy based models for anomaly detection. In *ICML*, 1100–1109.

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.

Zitnik, M.; and Leskovec, J. 2017. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, (14): i190–i198.