3.2. Energy-based Contrastive Learning

Theory Initially, we define p_d as the distribution of the graph data and $\mathcal{T}$ as a set of predetermined data augmentation operators. Given a dual-attribute subgraph $g^\dagger$ and two augmentation views $t, t' \sim \mathcal{T}$ selected uniformly at random, we propose an ECL approach to approximate the joint distribution $p_d(\nu, \nu')$ over two views $\nu = t(g^\dagger)$, $\nu' = t'(g^\dagger)$. The contrastive paradigm $f_\theta(\nu, \nu')$ is defined as:

$$p_\theta(\nu, \nu') = \frac{\exp(-f_\theta(\nu, \nu'))}{Z(\theta)}\,, \tag{5}$$

where $Z(\theta) = \int \exp(-f_\theta(\nu, \nu'))$, τ is a temperature parameter, and z are representations computed by feeding the augmentations ν into a GNN encoder $\varphi_\theta(\cdot)$ and a linear projection $\phi_\theta(\cdot)$ sequentially.

Building on the assumption that semantically similar pairs (ν, ν') will have nearby projections, corresponding to high p_d, while dissimilar ones will have distant projections, corresponding to low p_d, we solve for the distance between z and z' through $E_\theta(\cdot)$. The term $\|z - z'\|$ is used to indicate the inverse of semantic similarity of ν and ν'. To approximate $p_\theta(\nu, \nu')$ to $p_d(\nu, \nu')$, Eq. 1 can be rephrased as:

$$\min_\theta \mathbb{E}_{p_d}[-\log p_\theta(\nu, \nu')]\,. \tag{6}$$

Proposition 1. *The joint distribution $p_\theta(\nu, \nu')$ can be formulated as an EBM.*

$$p_\theta(\nu, \nu') = \frac{\exp(-E_\theta(\nu, \nu'))}{Z(\theta)}\,, \tag{7}$$

where $E_\theta(\nu, \nu') = \|z - z'\|^2/\tau$ and the derivative of the objective of Eq. 6 is expressed as:

$$\begin{aligned}\nabla_\theta \mathbb{E}_{p_d}[-\log p_\theta(\nu, \nu')] = \\ \mathbb{E}_{p_d}[\nabla_\theta E_\theta(\nu, \nu')] - \mathbb{E}_{p_\theta}[\nabla_\theta E_\theta(\nu, \nu')]\,.\end{aligned} \tag{8}$$

To avoid directly calculating $Z(\theta)$, we employ Bayes' rule (Bayes, 1763) to rewrite $\mathbb{E}_{p_d}[-\log p_\theta(\nu, \nu')]$ as follows:

$$\begin{aligned}\mathbb{E}_{p_d}[-\log p_\theta(\nu, \nu')] = \\ \mathbb{E}_{p_d}[-\log p_\theta(\nu'|\nu)] + \mathbb{E}_{p_d}[-\log p_\theta(\nu)]\,,\end{aligned} \tag{9}$$

where $p_\theta(\nu)$ is the marginal distribution of ν under $p_\theta(\nu, \nu')$.

Theorem 1. *The marginal distribution $p_\theta(\nu)$ is an EBM.*

$$p_\theta(\nu) = \frac{\exp(-E_\theta(\nu))}{Z(\theta)}\,, \tag{10}$$

where $E_\theta(\nu) = -\log \int e^{-\|z-z'\|^2/\tau} d\nu'$ and the gradient of the Eq. 10 is defined as:

$$\nabla_\theta \mathbb{E}_{p_d}[-\log p_\theta(\nu)] = \mathbb{E}_{p_d}[\nabla_\theta E_\theta(\nu)] - \mathbb{E}_{p_\theta}[\nabla_\theta E_\theta(\nu)]\,. \tag{11}$$

Here, the ECL objective is decomposed into the generative term and discriminative term, is given by:

$$\mathcal{L}_b(\theta) = \mathbb{E}_{p_d}[-\log p_\theta(\nu'|\nu)] + \alpha \mathbb{E}_{p_d}[-\log p_\theta(\nu)]\,, \tag{12}$$

where α is a hyperparameter to trade off the strength of two terms. According to Eq. 8, the gradient of the Eq. 12 is:

$$\begin{aligned}\nabla_\theta \mathcal{L}_b(\theta) = \mathbb{E}_{p_d}[-\nabla_\theta \log p_\theta(\nu'|\nu)] + \\ \alpha \mathbb{E}_{p_d}[\nabla_\theta E_\theta(\nu)] - \alpha \mathbb{E}_{p_\theta}[\nabla_\theta E_\theta(\nu)]\,.\end{aligned} \tag{13}$$

By adopting this way, $Z(\theta)$ ingeniously cancels itself out in the discriminative term without additional calculations. For the generative term, we merely need to sample ν^* from $p_d(\nu)$ with adding noise $\mathcal{N}(0, \lambda)$ and iteratively optimize ν^* through SGLD, as indicated in Eq. 3.

Implementation To commence the training of ECL, we approximate the two terms of Eq. 9, respectively, using the empirical mean of $p_\theta(\nu)$ (Kim & Ye, 2022). Given a mini-batch of samples $\{(\nu_n, \nu'_n)\}_{n=1}^N$ and their corresponding representations $\{(z_n, z'_n)\}_{n=1}^N$, the empirical mean $\hat{p}_\theta(\nu_n)$ is written as:

$$\hat{p}_\theta(\nu_n) = \frac{1}{2N} \sum_{\nu'_m : \nu'_m \neq \nu_n}^{2(N-1)} p_\theta(\nu_n, \nu'_m)\,. \tag{14}$$

For the discriminative term, we utilize $\frac{p_\theta(\nu_n, \nu'_n)}{\hat{p}_\theta(\nu_n)}$ to approximate the conditional probability density $p_\theta(\nu'|\nu)$. According the SimCLR framework (Chen et al., 2020a), $\min_\theta \mathbb{E}_{p_d}[-\log \hat{p}_\theta(\nu'_n|\nu_n)]$ can be represented as:

$$\min_{z \in f_\theta(\nu)} -\log\left(\frac{\exp\left(-\|z_n - z'_n\|^2/\tau\right)}{\frac{1}{2N}\sum_{\nu'_m : \nu'_m \neq \nu_n}^{2(N-1)} \exp\left(-\|z_n - z'_m\|^2/\tau\right)}\right). \tag{15}$$

Considering only N positive samples in the generative term, we simplify Eq. 14 to $\hat{p}_\theta(\nu_n) = \frac{1}{N}\sum_{m=1}^N p_\theta(\nu_n, \nu'_m)$. The approximation of $\min_\theta \mathbb{E}_{p_d}[-\log \hat{p}_\theta(\nu, \nu_n)]$ is denoted as:

$$\min_{z \in f_\theta(\nu)} -\log\left(\sum_{n=1}^N \exp\left(-\|z - z_n\|^2/\tau\right)\right). \tag{16}$$

In summation, the final objective of ECL is:

$$\mathcal{L}_E(\theta) = \mathcal{L}_b(\theta) + \beta \mathcal{L}_r(\theta)\,, \tag{17}$$

where $\mathcal{L}_r(\theta) = \frac{1}{2N}\sum_{m \neq n} E_\theta\left(\nu_n, \nu'_m\right)^2$ is L_2 regularization loss which serves to prevent gradient overflow due to the excessive energy values. β is a hyperparameter.

Table 1. Node classification accuracy (mean(%)±std) with the standard splits on various benchmark datasets. The top three results are highlighted in **first best**, **second best**, and **third best**, respectively. "OOM" indicates out of memory.

Method	Cora	Citeseer	Cornell	Texas	Wisconsin	Actor	Pubmed	Arxiv	Products	Proteins
GCN	81.46 ± 0.58	71.36 ± 0.31	47.84 ± 5.55	57.83 ± 2.76	57.45 ± 4.30	30.01 ± 0.77	**79.18** ± 0.29	**70.77** ± 0.19	**75.64**± 0.21	**72.51**± 0.35
GAT	81.41 ± 0.77	70.69 ± 0.58	46.22 ± 6.33	54.05 ± 7.35	57.65 ± 7.75	28.91 ± 0.83	77.85 ± 0.42	**69.90** ± 0.25	**79.45**± 0.59	**72.02**± 0.44
LDS	83.01 ± 0.41	**73.55** ± 0.54	47.87 ± 7.14	58.92 ± 4.32	61.70 ± 3.58	**31.05** ± 1.31	OOM	OOM	OOM	OOM
GEN	80.21 ± 1.72	71.15 ± 1.81	**57.02** ± 7.19	**65.94** ± 4.13	**66.07** ± 3.72	27.21 ± 2.05	78.91 ± 0.69	OOM	OOM	OOM
SGSR	83.48 ± 0.43	72.96 ± 0.25	44.32 ± 2.16	60.81 ± 4.87	56.86 ± 1.24	30.23 ± 0.38	78.09 ± 0.53	OOM	OOM	OOM
GRCN	**83.87** ± 0.49	72.43 ± 0.61	54.32 ± 8.24	62.16 ± 7.05	56.08 ± 7.19	29.97 ± 0.71	78.92 ± 0.39	OOM	OOM	OOM
IDGL	**83.88** ± 0.42	72.20 ± 1.18	50.00 ± 8.98	62.43 ± 6.09	59.41 ± 4.11	28.16 ± 1.41	OOM	OOM	OOM	OOM
GAuG-O	82.20 ± 0.80	71.60 ± 1.10	57.60 ± 3.80	56.90 ± 3.60	54.80 ± 5.70	25.80 ± 1.00	**79.30** ± 0.40	OOM	OOM	OOM
SUBLIME	83.40 ± 0.42	72.30 ± 1.09	**70.54** ± 5.98	**70.03** ± 4.23	**66.81** ± 6.55	30.79 ± 0.68	73.80 ± 0.60	55.50 ± 0.10	–	–
ProGNN	80.30 ± 0.57	68.51 ± 0.52	54.05 ± 6.16	48.37 ± 12.17	62.54 ± 7.56	22.35 ± 0.88	71.60 ± 0.46	OOM	OOM	OOM
CoGSL	81.76 ± 0.24	**73.09** ± 0.42	52.16 ± 3.21	59.46 ± 4.36	58.82 ± 1.52	**32.95** ± 1.20	OOM	OOM	OOM	OOM
STABLE	80.20 ± 0.68	68.91 ± 1.01	44.03 ± 4.05	55.24 ± 6.04	53.00 ± 5.27	30.18 ± 1.00	OOM	OOM	OOM	OOM
NodeFormer	80.28 ± 0.82	71.31 ± 0.98	42.70 ± 5.51	58.92 ± 4.32	48.43 ± 7.02	25.51 ± 1.17	78.21 ± 1.43	55.40 ± 0.23	–	–
ECL-GSR	**84.06** ± 0.84	**73.70** ± 0.75	**71.27** ± 2.06	**72.97** ± 3.39	**67.79** ± 1.03	**33.71** ± 0.96	**80.91** ± 1.12	**71.09** ± 0.31	**80.47**± 0.22	**74.64**± 0.31

Table 2. Pairwise independent sample t-tests comparison of ECL-GSR with other methods on Cora, Citeseer, Actor, and Pubmed datasets.

Method	Cora		Citeseer		Actor		Pubmed	
	T-statistic	*P-value*	*T-statistic*	*P-value*	*T-statistic*	*P-value*	*T-statistic*	*P-value*
GCN	12.33	4.24×10^{-9}	7.85	4.18×10^{-6}	11.45	1.40×10^{-8}	5.46	3.66×10^{-3}
GAT	11.15	2.12×10^{-8}	11.00	2.62×10^{-8}	14.54	2.83×10^{-10}	6.89	2.51×10^{-5}
LDS	6.41	6.37×10^{-5}	11.62	8.97×10^{-7}	10.14	3.96×10^{-3}	–	–
GEN	5.66	2.97×10^{-4}	8.22	4.65×10^{-4}	14.72	2.31×10^{-10}	8.51	3.22×10^{-4}
SGSR	10.21	6.84×10^{-4}	5.24	1.55×10^{-2}	13.35	1.15×10^{-9}	4.71	2.29×10^{-3}
GRCN	6.95	1.12×10^{-3}	13.06	2.04×10^{-5}	11.71	9.69×10^{-9}	4.21	6.86×10^{-3}
IDGL	7.58	2.47×10^{-3}	3.99	8.05×10^{-2}	13.40	1.08×10^{-9}	–	–
GAuG-O	5.92	1.71×10^{-4}	4.22	6.67×10^{-3}	16.83	2.42×10^{-11}	9.10	8.02×10^{-3}
SUBLIME	8.09	8.85×10^{-4}	7.49	4.86×10^{-4}	8.33	1.77×10^{-6}	16.04	5.45×10^{-11}
ProGNN	16.33	3.99×10^{-11}	18.10	6.97×10^{-12}	25.95	1.34×10^{-14}	21.63	3.22×10^{-13}
CoGSL	13.06	1.66×10^{-9}	4.67	3.67×10^{-2}	12.37	3.81×10^{-5}	–	–
STABLE	15.52	9.45×10^{-11}	9.43	2.86×10^{-7}	7.58	6.86×10^{-6}	–	–
NodeFormer	12.26	4.62×10^{-9}	6.09	1.22×10^{-4}	14.36	3.46×10^{-10}	4.90	1.49×10^{-3}

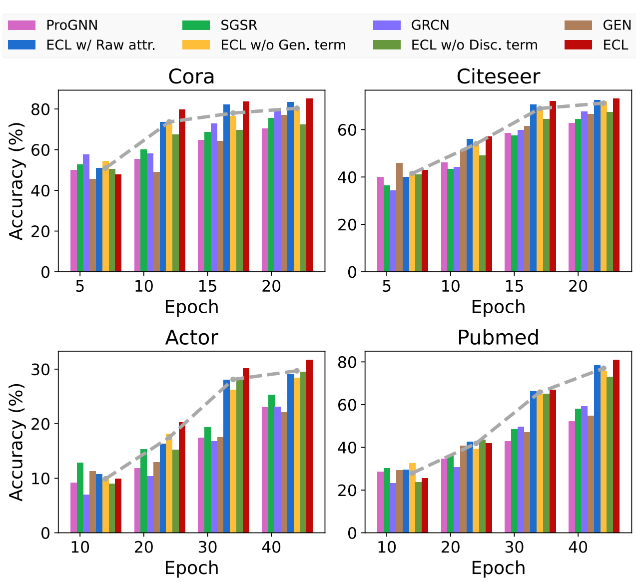

Figure 1. Performance analysis of ECL-GSR variants and other benchmarks over a range of training epochs on four datasets.

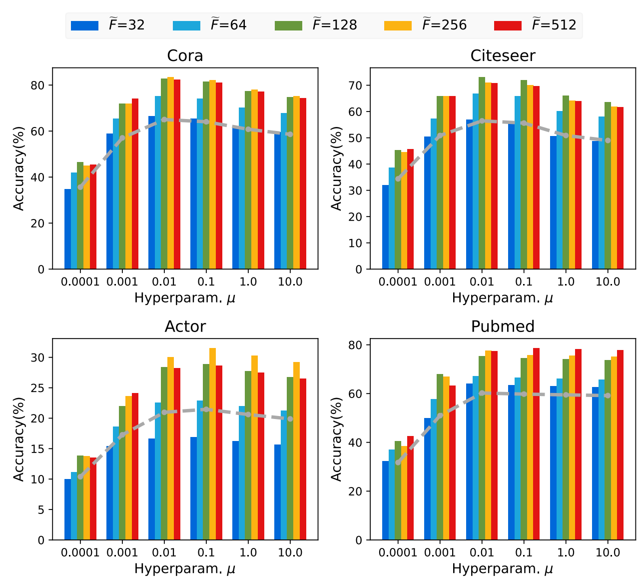

Figure 2. Hyperparameter μ and dimensionality $\widetilde{F}$ analysis of ECL-GSR on Cora, Citeseer, Actor, and Pubmed datasets.