The 33rd International Joint Conference on Artificial Intelligence

Rethinking the Effectiveness of Graph Classification Datasets in Benchmarks for Assessing GNNs

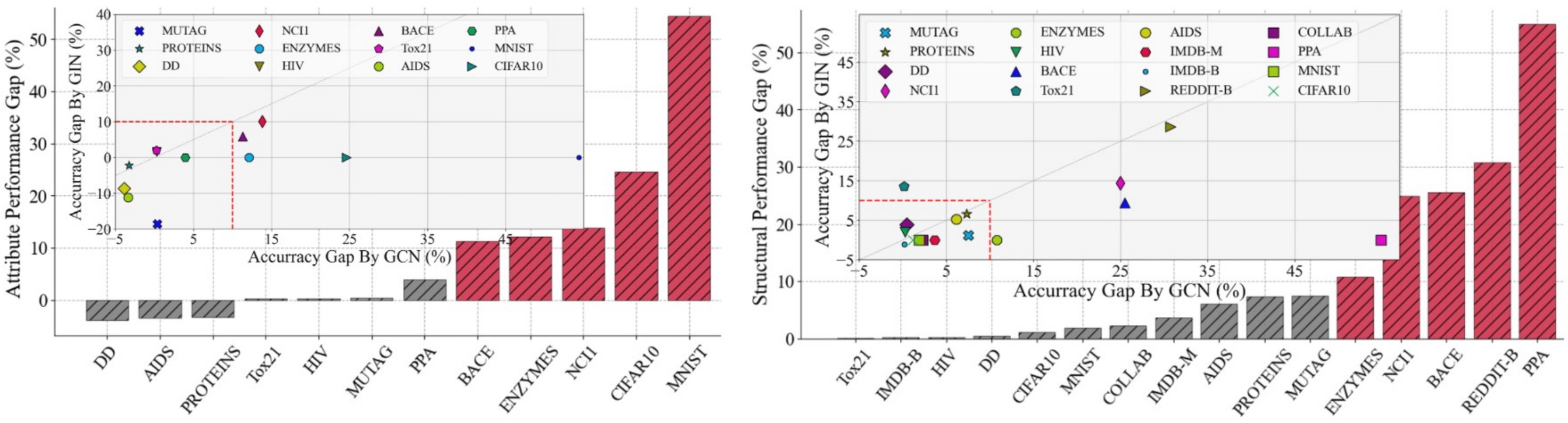
Zhengdao Li[1,2], Yong Cao[3], Kefan Shuai[1], Yiming Miao[1*] and Kai Hwang[1]
[1]The Chinese University of Hong Kong, Shenzhen, China
[2]Guangzhou University, Guangzhou, China
[3]Huazhong University of Science and Technology, Wuhan, China
{zhengdaoli, kefanshuai}@link.cuhk.edu.cn
yongcao2018@gmail.com
{miaoyiming, hwangkai}@cuhk.edu.cn

Research Questions:

RQ1: Can commonly used graph classification datasets serve the benchmarking purpose which is to effectively distinguish advancements of GNNs compared with other methods?

RQ2: How to measure the effectiveness of existing graph classification datasets?

Observations:



(a) Attributed accuracy gap (b) Structural accuracy gap

Figure 1: The performance gaps on 16 graph classification datasets are categorized into two types: *Ineffective* (gray) and *Effective* (red) benchmarks. These are sorted in ascending order based on the size of the performance gap. An empirical threshold of 10% is used for categorization, as observed in the inner box of each figure. This box represents the distribution of the accuracy gap for GCN and GIN.

Definition1 ---- Performance gap

Given a dataset D, a baseline method $\mathcal{M}_{\textbf{type}}^{\text{Baseline}}$, and a graph-based method $\mathcal{M}_{\textbf{type}}^{\text{Graph}}$, the performance gap $\delta_{\textbf{type}}(D, R, \mathcal{M}^{\text{Graph}}, \mathcal{M}^{\text{Baseline}})$ (simply denoted by $\delta_{\textbf{type}}$) between baseline and graph-based method is defined as:

$$\delta_{\textbf{type}} \triangleq R(D, \mathcal{M}_{\textbf{type}}^{\text{Graph}}) - R(D, \mathcal{M}_{\textbf{type}}^{\text{Baseline}}),\ \textbf{type} \in \{\mathbf{S}, \mathbf{A}\},$$

Limitations of the performance gap:

$$R\left(D_1, \mathcal{M}^{GNN}\right) = 100\%$$

= 10%

$$R\left(D_1, \mathcal{M}^{Baseline}\right) = 90\%$$

$$R(D_2, \mathcal{M}^{GNN}) = 60\%$$

= 10%

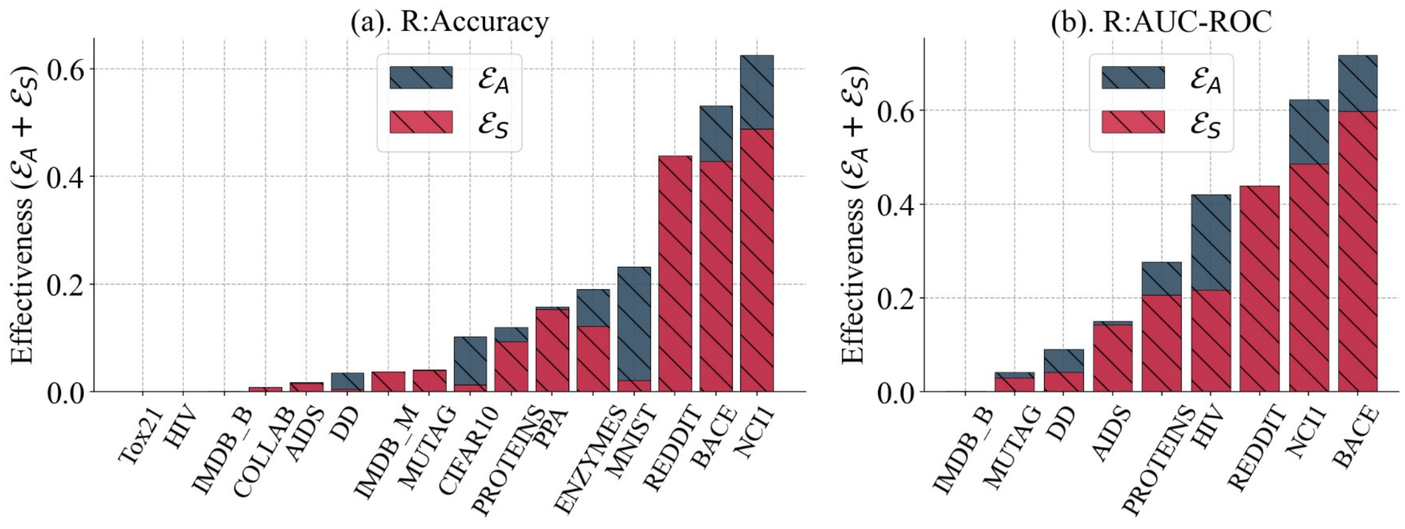
$$R(D_2, \mathcal{M}^{Baseline}) = 50\%$$

Proposed metric:

Definition 2 ---- Dataset Effectiveness

Given a graph classification dataset D which has $|Y|$ classes, and the performance gap $\delta_{\text{type}}(D)$ between two methods $\mathcal{M}^1$ and $\mathcal{M}^2$, the $\mathcal{E}$ to quantify the discriminating degree of $\mathcal{M}^1$ and $\mathcal{M}^2$ is defined as follows:

$$\mathcal{E}(D) = \sum_{\text{type}\in\{\mathbf{S},\mathbf{A}\}} \frac{|\delta_{\text{type}}(D)|}{R^*(|Y|-1)} \cdot \frac{1-R^*}{1-|Y|^{-1}},$$



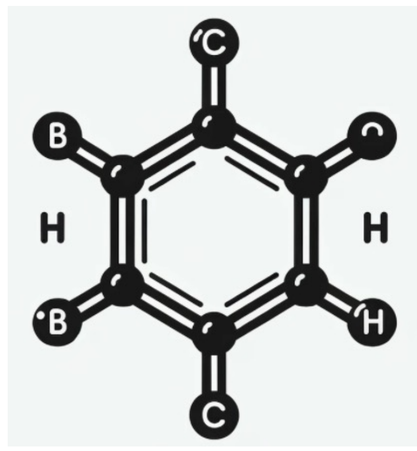


Figure 3: Effectiveness using Accuracy metric and AUC-ROC metric in terms of structural type and attributed type.

Investigation of low effectiveness

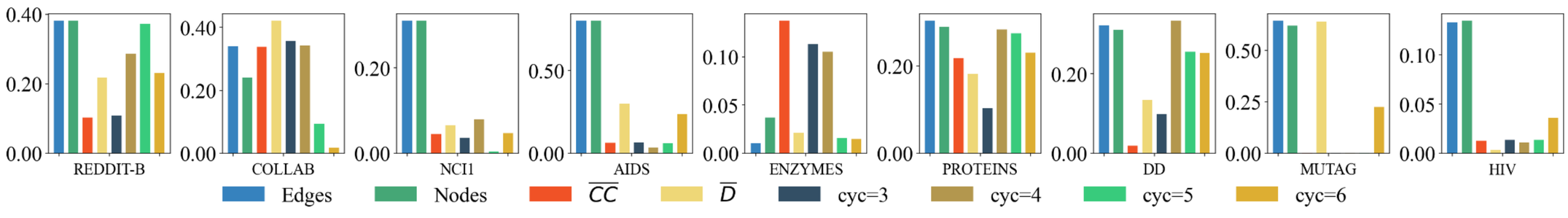


Figure 4: Correlations between graph property sequences and class labels on 9 real-world datasets.

Synthetic dataset generation

Theorem 1:

Given a set of property variables $\{\mathcal{P}_i\}_{i=1}^K$, *each* $\mathcal{P}_i$ *follows a Gaussian distribution* $\mathcal{N}(\mu_k, \sigma_k)$ *or Uniform distribution* $\mathcal{U}(a_k, b_k)$, *and given corresponding Pearson correlation coefficients* $\{r_i\}_{i=1}^K$ *with label variable* $\mathcal{Y}$, *with the constraint* $\sum_{i=1}^K r_i^2 \leq 1$, then we have:

$$\mathcal{Y} = \sigma_{\mathcal{Y}} \left(\sum_{i=1}^{K} n_i r_i + n_0 \sqrt{1 - \sum_{i=1}^{K} r_i^2} \right), \tag{1}$$

where $\sigma_{\mathcal{Y}}$ *is any desired standard deviation, and each* n_i *is mutually independent and follows the same distribution as the corresponding* $\mathcal{P}_i$ *with the same mean value* μ_i *but with standard deviation equals to 1.* (The proof is based on Cholesky decomposition of a given covariance matrix.)

Algorithm 1: Controllable dataset construction

Input: $\{r_k\}_{k=1}^K$, number of labels C,
$\{\mathcal{P}_k\}_{k=1}^K \sim \mathcal{N}(\mu_k, \sigma_k)$ or $\mathcal{U}(a_k, b_k)$;
Output: Dataset $\mathbb{D}$ with size N;
for $k = 1$ **to** K **do**
  Sample $n_k \sim \mathcal{N}(0, \sigma_k)$ or $\mathcal{U}(-\sqrt{3}, \sqrt{3})$;
  $\mathbb{P}_k \leftarrow \mu_k + \sigma_k n_k$ or $\frac{a_k+b_k}{2} + \sqrt{\frac{b_k-a_k}{12}} n_k$;
end
Calculate $\mathcal{Y}$ by the Eq. 2 ;
$\mathbb{Y} \leftarrow \text{ROUND}(\text{NORM}(\mathcal{Y}) * C)$;
$\mathbb{D} \leftarrow \{(g_i, y_i)\}_{i=1}^N$,
where each graph g_i has properties $\{\mathbb{P}_k[i]\}_{k=1}^K$, and corresponding label $y_i = \mathbb{Y}[i]$;

Prediction of Effectiveness

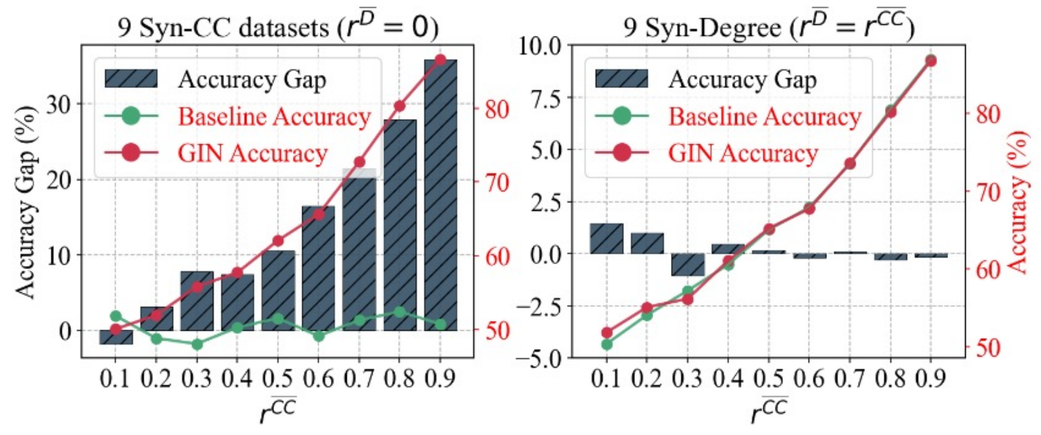


Figure 6: Controllable performance gaps by two types of synthetic datasets.

Regressor	Real-world datasets		Synthetic-CC datasets	
	Pearson	P-Value	Pearson	P-Value
Ridge	0.80 ± 0.09	$\leq 1 \times 10^{-6}$	0.87 ± 0.03	$\leq 1 \times 10^{-6}$
SVR	0.80 ± 0.09	$\leq 1 \times 10^{-6}$	0.89 ± 0.04	$\leq 1 \times 10^{-6}$
RF	0.89 ± 0.03	$\leq 1 \times 10^{-6}$	0.87 ± 0.06	$\leq 1 \times 10^{-6}$

Table 3: Summary of regression results