Graph Contrastive Learning Automated
(Appendix)

Yuning You [1] **Tianlong Chen** [2] **Yang Shen** [1] **Zhangyang Wang** [2]

A. Alternating Gradient Descent for JOAOv2

We adapt alternating gradient descent (AGD) in Algorithm 1 to optimize (10) in main text, executed as Algorithm S1, with the following modified upper-level minimization and lower-level maximization.

Algorithm S1 AGD for optimization (10) in main text

Input: initial parameter $\theta^{(0)}$, sampling distribution $\mathbb{P}^{(0)}_{(\mathsf{A}_1,\mathsf{A}_2)}$, $\mathbb{P}^{(0)}_{(\Theta_1'',\Theta_2'')} = \mathbb{P}^{(0)}_{(\mathsf{A}_1,\mathsf{A}_2)}$, optimization step N.
for $n = 1$ **to** N **do**
1. Upper-level minimization: fix $\mathbb{P}^{(n-1)}_{(\mathsf{A}_1,\mathsf{A}_2)}, \mathbb{P}^{(n-1)}_{(\Theta_1'',\Theta_2'')}$, and call equation (1) to update $\theta^{(n)}$.
2. Lower-level maximization: fix $\theta^{(n)}$, call equation (3) to update $\mathbb{P}^{(n)}_{(\mathsf{A}_1,\mathsf{A}_2)}$, and set $\mathbb{P}^{(n)}_{(\Theta_1'',\Theta_2'')} = \mathbb{P}^{(n)}_{(\mathsf{A}_1,\mathsf{A}_2)}$.

end for
Return: Optimized parameter $\theta^{(N)}$.

Upper-level minimization. The upper-level minimization w.r.t. θ given the sampling distribution $\mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)}$, setting $\mathbb{P}_{(\Theta_1'',\Theta_2'')} = \mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)}$, is represented as:

$$\theta^{(n)} = \theta^{(n-1)} - \alpha' \nabla_\theta \mathcal{L}_{\text{v2}}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta', \Theta_1'', \Theta_2''), \quad (1)$$

where $\alpha' \in \mathcal{R}_{>0}$ is the learning rate.

Lower-level maximization. To calculate the gradient of the lower-level objective w.r.t. $\mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)}$, we make similar efforts to approximate the contrastive loss as:

$$\begin{aligned}
&\mathcal{L}_{\text{v2}}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta', \Theta_1'', \Theta_2'') \\
\approx &\sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|} p_{ij}\ell_{\text{v2}}(\mathsf{G}, A^i, A^j, \theta', \theta''^i, \theta''^j) \\
= &\sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|} p_{ij}\Big\{ - \mathbb{E}_{\mathbb{P}_\mathsf{G}}\text{sim}(T^i_{\text{v2},\theta}(\mathsf{G}), T^j_{\text{v2},\theta}(\mathsf{G})) \\
&+ \mathbb{E}_{\mathbb{P}_\mathsf{G}}\log(\mathbb{E}_{\mathbb{P}_{\mathsf{G}'}}\exp(\text{sim}(T^i_{\text{v2},\theta}(\mathsf{G}), T^j_{\text{v2},\theta}(\mathsf{G}'))))\Big\}, \quad (2)
\end{aligned}$$

where $T^i_{\text{v2},\theta} = A^i \circ f_{\theta'} \circ g_{\theta''^i},\ i = 1, ..., 5$. Thus, projected gradient descent is performed as:

$$\boldsymbol{b} = \boldsymbol{p}^{(n-1)} + \alpha'' \nabla_{\boldsymbol{p}} \psi_{\text{v2}}(\boldsymbol{p}^{(n-1)}), \boldsymbol{p}^{(n)} = (\boldsymbol{b} - \mu\mathbf{1})_+, \quad (3)$$

where $\boldsymbol{p} = [p_{ij}], i, j = 1, ..., |\mathcal{A}|, \psi_{\text{v2}}(\boldsymbol{p}) = \sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|} p_{ij}\ell_{\text{v2}}(\mathsf{G}, A^i, A^j, \theta', \theta''^i, \theta''^j) - \frac{\gamma}{2}\sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|}(p_{ij} - \frac{1}{|\mathcal{A}|^2})^2$, $\alpha'' \in \mathcal{R}_{>0}$ is the learning rate, μ is the root of the equation $\mathbf{1}^\mathsf{T}(\boldsymbol{b} - \mu\mathbf{1}) = 1$, and $(\cdot)_+$ is the element-wise non-negative operator.

B. Dataset Statistics

Dataset statistics can be found in Table S1, S2 and S3.

Table S1: Statistics for datasets of diverse nature from the benchmark TUDataset.

Dataset	Graph Count	Avg. Node	Avg. Degree
NCI1	4,110	29.87	1.08
PROTEINS	1,113	39.06	1.86
DD	1,178	284.32	715.66
MUTAG	188	17.93	19.79
COLLAB	5,000	74.49	32.99
RDT-B	2,000	429.63	1.15
RDB-M	2,000	429.63	497.75
GITHUB	4,999	508.52	594.87
IMDB-B	1,000	19.77	96.53

[1]Texas A&M University [2]The University of Texas at Austin. Correspondence to: Yang Shen <yshen@tamu.edu>, Zhangyang Wang <atlaswang@utexas.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

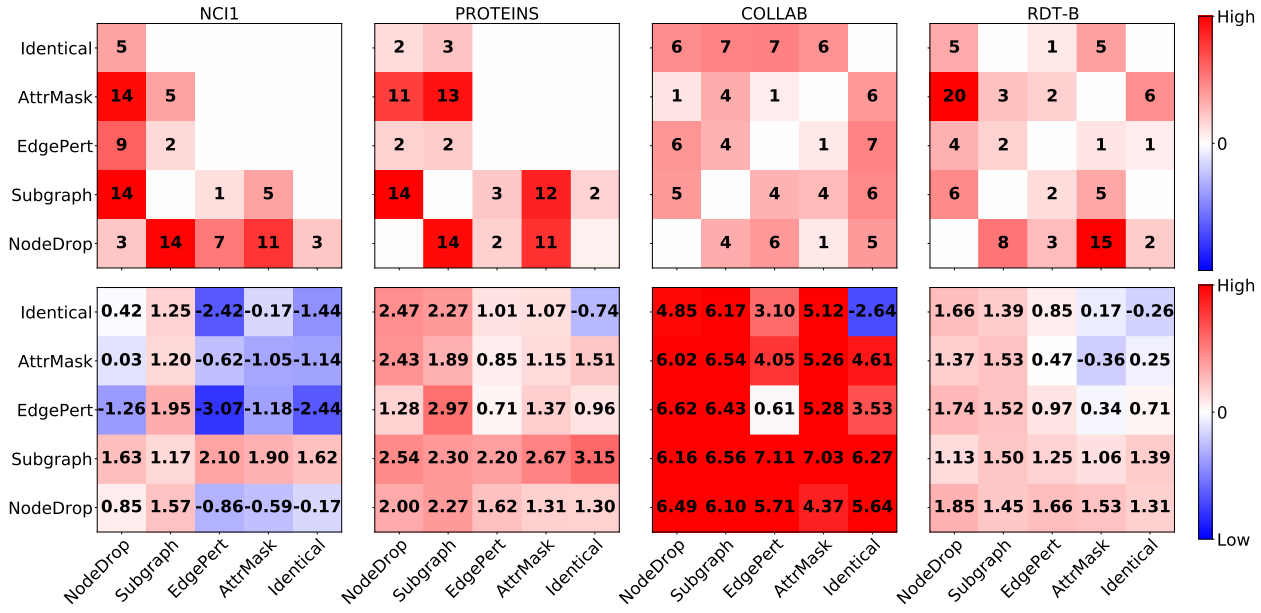

Figure S1: Top row: sampling distributions (%, defined as the percentage of this specific augmentation pair being selected during the entire training process) for augmentation pairs selected by JOAOv2 on four different datasets (NCI1, PROTEINS, COLLAB, and RDT-B). **Bottom row**: GraphCL performance gains when exhaustively trying every possible augmentation pair. Warmer (colder) colors indicate higher (lower) values, and white marks 0.

Table S2: Statistics for bioinformatics datasets.

Dataset	Graph Count	Avg. Node	Avg. Degree
BBBP	2,039	24.06	51.90
Tox21	7,831	18.57	38.58
ToxCast	8,576	18.78	38.52
SIDER	1,427	33.64	70.71
ClinTox	1,477	26.15	55.76
MUV	93,087	24.23	52.55
HIV	41,127	25.51	54.93
BACE	1,513	34.08	73.71
PPI	88,000	49.35	890.77

Table S3: Statistics for large-scale OGB datasets.

Dataset	Graph Count	Avg. Node	Avg. Degree
ogbg-ppa	158,110	243.4	2,266.1
ogbg-code	452,741	125.2	124.2

C. Augmentation Sampling Rules for GraphCL

GraphCL uniformly samples augmentations from a pre-defined pool. Augmentation pools for datasets are presented in Table S4.

Table S4: Pre-defined augmentation pools for different datasets.

Datasets	Augmentation Pools
NCI1	{NodeDrop, Subgraph}
PROTEINS	{NodeDrop, Subgraph}
DD	{NodeDrop, Subgraph}
MUTAG	{NodeDrop, Subgraph}
COLLAB	{NodeDrop, Subgraph, EdgePert, AttrMask}
RDT-B	{NodeDrop, Subgraph, EdgePert}
RDB-M	{NodeDrop, Subgraph, EdgePert}
GITHUB	{NodeDrop, Subgraph, EdgePert, AttrMask}
IMDB-B	{NodeDrop, Subgraph}
BBBP	{NodeDrop, Subgraph}
Tox21	{NodeDrop, Subgraph}
ToxCast	{NodeDrop, Subgraph}
SIDER	{NodeDrop, Subgraph}
ClinTox	{NodeDrop, Subgraph}
MUV	{NodeDrop, Subgraph}
HIV	{NodeDrop, Subgraph}
BACE	{NodeDrop, Subgraph}
PPI	{NodeDrop, Subgraph, EdgePert}
ogbg-ppa	{NodeDrop, Subgraph, EdgePert}
ogbg-code	{NodeDrop, Subgraph}

D. JOAOv2 Selected Augmentation-Pairs Alignment with Previous "Best Practices"

Schematic diagram of JOAOv2 is drawn in Figure S2. $\mathbb{P}_{(A_1,A_2)}$ optimized by JOAOv2 is plotted in the top row of Figure S1 along with GraphCL performance gains of different augmentation pairs in the bottom row, following

the same procedure as described in Sec. 3.2.1 of main text.

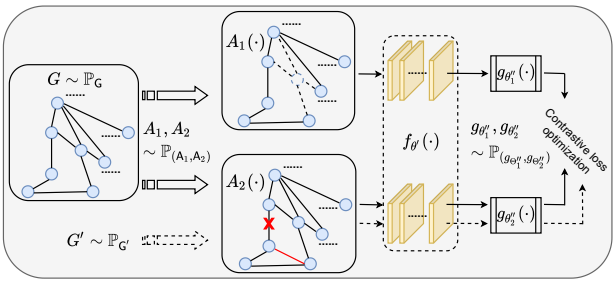

Figure S2: Schematic diagram of GraphCL with multiple augmentation-aware projection heads where $\mathbb{P}_{(g_{\Theta_1''}, g_{\Theta_2''})} = \mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)}$.

E. Comparison between JOAO w/ and w/o the Prior

See Table S5 for comparison between JOAO w/ and w/ the prior in the semi-supervised learning setting.

Table S5: Semi-supervised performance (%) of JOAO w/ and w/o prior.

γ	w/o prior	w/ prior
NCI1	61.51±0.32	61.97±0.72
PROTEINS	71.78±0.70	72.13±0.92