Response to Reviewers

Ziyang Liu[1], Chaokun Wang[1], Liqun Yang[2], Yunkai Lou[1], Hao Feng[1], Cheng Wu[1], Kai Zheng[3], Yang Song[3]
[1]*Tsinghua University*, Beijing, China; [2]*Nanhu Academy*, Jiaxing, China; [3]*Kuaishou Technology*, Beijing, China
{liu-zy21,louyk18,fh20,wuc22}@mails.tsinghua.edu.cn, chaokun@tsinghua.edu.cn,
yangliqun@cnaeit.com, zhengkai@kuaishou.com, ys@sonyis.me

Firstly, we would like to thank the meta-reviewer and reviewers for their insight and valuable comments to our ICDE 2024 submission (661). We have carefully taken your comments into consideration in preparing this revision. Below we briefly summarize how we have specifically addressed the comments of each reviewer in our revised manuscript. We believe that our response has sufficiently addressed all the issues raised by the reviewers. In our revised paper, all the modifications are marked in blue. In the responses, the deleted, updated, and added texts are marked by ~~strike-through~~, orange, and blue, respectively.

I. Response to Meta-Reviewer

A. Area Chair's Summary

Comment: The reviewers found some merits of this submission. However, some concerns are also raised. We are happy to invite the authors to revise the paper and address all the comments and concerns from the reviewers.

Response Thanks for your kind suggestion. We have replied to all the comments from the reviewers and finished all the required experiments suggested by the reviewers:

- We have addressed two reviewers' confusion regarding the connection between this paper and data mining problems, and cited more papers from database conferences. (Reviewer #1 O1, Reviewer #3 O6).
- We have incorporated the method with the same τ setting as described in the original paper of GRACE into our ablation studies. (Reviewer #2 O1).
- We have deployed GLATE with more baseline methods (DGI, SUGRL, and NCLA) in the inductive node classification task to verify its effectiveness (Reviewer #2 O2).
- We have conducted the sensitivity analysis to more hyperparameters in GLATE, including the initial momentum, momentum parameter, similarity threshold, temperature learning rate, and sampled node count (Reviewer #2 O3).
- We have reviewed the entire paper and addressed the minor issues with the figures, e.g., Fig. 4. (Reviewer #2 O4).
- We have added more efficiency analysis of GLATE to demonstrate the high efficiency of temperature fine-tuning. Also, we have completed the sensitivity analysis experiments on different types of datasets to show that a general parameter configuration can achieve satisfactory performance when training GLATE on a new graph dataset without excessive parameter tuning time. (Reviewer #3 O1).
- We have evaluated our method and demonstrated its versatility by varying the density, connectivity patterns, and node/edge attributes in graphs (Reviewer #3 O2).
- We have added the scalability analysis to demonstrate that GRACE+GLATE is scalable to large-scale graphs. The experimental results on time costs and memory peak are consistent with our complexity analysis. (Reviewer #3 O3).
- We have applied our method to more downstream tasks including link prediction, graph classification, and node clustering, demonstrating its potential for diverse graph-related tasks in real-world applications. (Reviewer #3 O4).
- We have added the case study to interpret the learned embeddings in molecule graphs. The analysis indicates that the embeddings learned by GRACE+GLATE align with fundamental chemical knowledge well. (Reviewer #3 O5).

II. Response to Reviewer #1

A. O1

Comment: The paper is a machine learning paper. The main body of the paper is about how to improve the training process of graph contrastive learning. It is not specific for a data engineering or data management/mining problem. This can also be seen from the references of the paper. The majority of the cited papers are from machine learning venues. As such, it would be better for the paper to be evaluated in a machine learning venue.

Response: Thanks for your comment. In this paper, we focus on improving the training process of graph contrastive learning (GCL). It is important to note that GCL supports downstream tasks, which are integral to the field of data mining. For example, node classification is important for social/citation network analysis; graph classification is important for molecule property prediction; link prediction is important for recommender systems. In the revised manuscript, we apply GLATE to four different downstream tasks, including node classification, graph classification, node clustering, and link prediction. Our proposed method GLATE not only improves the performance of the base model but also accelerates its training process, which makes GCL models easily scalable to large-scale graph data.

Furthermore, we have investigated the works closely related to ours that were published in SIGMOD, VLDB, and ICDE, and added 16 new references into the revised manuscript. These papers cover self-supervised learning ([1], [2], [3], [21], and [22]), node classification or link prediction on temporal graphs ([29], [31], [32], [38] and [39]), citation network

analysis ([30], [33], and [36]), molecule property prediction ([34] and [35]), and link prediction on knowledge graphs ([37]).

The main revisions for the citations in the revised manuscript are as follows:

Raw texts: Self-supervised learning provides a promising learning paradigm without relying on high-cost label information for many research fields such as computer vision [1]–[3], natural language processing [4]–[7], speech recognition [8]–[10], and recommender systems [11]–[13]. Contrastive-based methods have a prominent place among the landscape of self-supervised learning methods [14]–[17]. Contrastive learning leverages the inherent structure and relationships within unlabeled data to train encoder networks [18]–[20]. ... On benchmark datasets, the state-of-the-art GCL models have demonstrated competitive performance against supervised learning models, e.g., graph convolutional network (GCN) [26], in various graph-related tasks such as node classification [23], [27] and graph classification [28], [29]

Revised texts: Self-supervised learning [1-3] provides a promising learning paradigm without relying on high-cost label information for many research fields such as computer vision [4]–[6], natural language processing [7]–[10], speech recognition [11]–[13], and recommender systems [14]–[16]. Contrastive-based methods have a prominent place among the landscape of self-supervised learning methods [17]–[20]. Contrastive learning leverages the inherent structure and relationships within unlabeled data to train encoder networks [21], [22]. ... On benchmark datasets, the state-of-the-art GCL models have demonstrated competitive performance against supervised learning models, e.g., graph convolutional network (GCN) [28], in various graph-related tasks such as node classification [29-33], graph classification [34], [35], and link prediction [36-39].

Specifically, the newly added references include:

[1] S. Guo, Y. Lin, L. Gong, C. Wang, Z. Zhou, Z. Shen, Y. Huang, and H. Wan, "Self-supervised spatial-temporal bottleneck attentive network for efficient long-term traffic forecasting," in 39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023, pp. 1585–1596, IEEE, 2023.

[2] R. Wang, Y. Li, and J. Wang, "Sudowoodo: Contrastive self-supervised learning for multi-purpose data integration and preparation," in 39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023, pp. 1502–1515, IEEE, 2023.

[3] J. Peng, D. Shen, N. Tang, T. Liu, Y. Kou, T. Nie, H. Cui, and G. Yu, "Self-supervised and interpretable data cleaning with sequence generative adversarial networks," Proc. VLDB Endow., vol. 16, no. 3, pp. 433–446, 2022.

[21] Y. Chang, J. Qi, Y. Liang, and E. Tanin, "Contrastive trajectory similarity learning with dual-feature attention," in 39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023, pp. 2933–2945, IEEE, 2023.

[22] W. Wang, B. Hu, Z. Peng, M. Zhong, Z. Zhang, Z. Liu, G. Zhang, and J. Zhou, "GARCIA: powering representations of long-tail query with multi-granularity contrastive learning," in 39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023, pp. 3182–3195, IEEE, 2023.

[29] J. Layne, J. Carpenter, E. Serra, and F. Gullo, "Temporal sir-gn: Efficient and effective structural representation learning for temporal graphs," Proceedings of the VLDB Endowment, vol. 16, no. 9, pp. 2075–2089, 2023.

[30] Y. Zhang and A. Kumar, "Lotan: Bridging the gap between gnns and scalable graph analytics engines," Proceedings of the VLDB Endowment, vol. 16, no. 11, pp. 2728–2741, 2023.

[31] F. Xiao, Y. Wu, M. Zhang, G. Chen, and B. C. Ooi, "Mint: Detecting fraudulent behaviors from time-series relational data," Proceedings of the VLDB Endowment, vol. 16, no. 12, pp. 3610–3623, 2023.

[32] X. Du, X. Zhang, S. Wang, and Z. Huang, "Efficient tree-svd for subset node embedding over large dynamic graphs," Proceedings of the ACM on Management of Data, vol. 1, no. 1, pp. 1–26, 2023.

[33] X. Zhang, Y. Shen, Y. Shao, and L. Chen, "Ducati: A dual-cache training system for graph neural networks on giant graphs with the gpu," Proceedings of the ACM on Management of Data, vol. 1, no. 2, pp. 1–24, 2023.

[34] G. Lv and L. Chen, "On data-aware global explainability of graph neural networks," Proceedings of the VLDB Endowment, vol. 16, no. 11, pp. 3447–3460, 2023.

[35] G. Lv, C. J. Zhang, and L. Chen, "Hence-x: Toward heterogeneity-agnostic multi-level explainability for deep graph networks," Proceedings of the VLDB Endowment, vol. 16, no. 11, pp. 2990–3003, 2023.

[36] C. Yang and J. Han, "Revisiting citation prediction with cluster-aware text-enhanced heterogeneous graph neural networks," in 39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023, pp. 682–695, IEEE, 2023.

[37] Y. Zhang, W. Wang, H. Yin, P. Zhao, W. Chen, and L. Zhao, "Disconnected emerging knowledge graph oriented inductive link prediction," in 39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023, pp. 381–393, IEEE, 2023.

[38] Y. Li, Y. Shen, L. Chen, and M. Yuan, "Orca: Scalable temporal graph neural network training with theoretical guarantees," Proceedings of the ACM on Management of Data, vol. 1, no. 1, pp. 1–27, 2023.

[39] H. Li and L. Chen, "Early: Efficient and reliable graph neural network for dynamic graphs," Proceedings of the ACM on Management of Data, vol. 1, no. 2, pp. 1–28, 2023.

TABLE I: Accuracy results (%) under different strategies.

Model	Cora	CiteSeer	PubMed
GRACE+GLATE	**84.8±0.3**	**74.4±0.1**	**87.2±0.1**
w/ hard negative sampling	83.1±0.4	67.1±0.6	85.3±0.1
w/ static τ (τ=0.8)	*84.3±0.1*	67.8±0.2	85.2±0.1
w/ static τ (τ=0.2)	83.1±0.3	64.0±0.5	82.1±0.1
w/ static τ (τ=0.4/0.9/0.7)	80.6±0.2	71.1±0.7	83.5±0.0
w/o τ_{lower} [†]	18.4±0.1	12.6±0.2	25.6±0.0
w/o momentum	83.7±0.2	*72.7±0.6*	*86.9±0.1*
TaU	83.5±0.2	67.8±0.4	85.5±0.0
iSogCLR	83.2±0.1	66.3±0.4	84.0±0.1
Adap-τ	83.6±0.2	68.5±0.3	85.9±0.1

[†] The value of training loss eventually becomes "NaN".

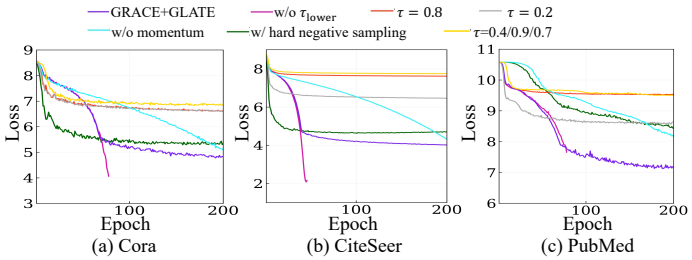


Fig. 1: Training loss curves of GRACE+GLATE and its variants.

III. Response to Reviewer #2

A. O1

Comment: In the ablation study, the static tau is not set the same as it is in the GRACE. In GRACE, tau is set to different values on these three datasets. Therefore, it is advised to use the same parameter settings as in the original paper.

Response: Thanks for your suggestion. We have added an experiment to use the same parameter settings for τ as reported in the original paper of GRACE. We denote this method as "w/ static τ (τ=0.4/0.9/0.7)", which means that fixing τ as the same value in GRACE (i.e., 0.4 for Cora, 0.9 for CiteSeer, and 0.7 for PubMed). We have updated this method's accuracy results and loss curve (yellow curve) in Table I (i.e., Table XI in the revised manuscript) and Fig. 1 (i.e., Fig. 2 in the revised manuscript), respectively. The updated results still support our original experimental conclusion, i.e., GRACE+GLATE has obvious advantages over its variants with static τ either on accuracy performance or training loss.

The main revisions for the experimental analysis of ablation studies are shown as follows:

Raw texts: **(ii) w/ static** τ: removing dynamic temperature estimation and fixing τ as 0.8 or 0.2; **(iii)** w/o τ_{lower}: ... We conduct ablation studies to assess the effects of different temperature strategies concerning negatives, temperature, or momentum on model performance. ... As shown in Table VIII, by estimating temperatures dynamically, GRACE+GLATE outperforms two models with static temperatures ($\tau = 0.8$ or $\tau = 0.2$), especially on CiteSeer. ... The results demonstrate that "$\tau = 0.2$", "$\tau = 0.8$", "w/o momentum", and GRACE+GLATE all converge rapidly, and our GRACE+GLATE model converges to the lowest loss value among them.

TABLE II: Accuracy results (%) on inductive node classification.

Model	PPI	Model	PPI
Raw model	42.2±0.0	DGI	63.8±0.2
GraphSAGE-GCN	46.5±0.2	DGI+GLATE (ours)	67.0±0.0
GraphSAGE-Mean	48.6±0.1	SUGRL	66.0±0.1
GraphSAGE-LSTM	48.2±0.1	SUGRL+GLATE (ours)	67.8±0.1
GraphSAGE-Pool	50.2±0.2	NCLA	66.2±0.1
GMI	65.0±0.0	NCLA+GLATE (ours)	*68.0±0.0*
GCA	66.0±0.1	GRACE	66.2±0.1
BGRL	67.5±0.1	GRACE+GLATE (ours)	**68.4±0.0** (p=8.1e-11)

Revised texts: **(ii) w/ static** τ: removing dynamic temperature estimation and fixing τ as 0.8 or 0.2; **(iii) w/ static** τ **(τ=0.4/0.9/0.7)**: fixing τ as the same value in GRACE (i.e., 0.4 for Cora, 0.9 for CiteSeer, and 0.7 for PubMed); **(iv)** w/o τ_{lower}: ... We conduct ablation studies to assess the effects of different temperature strategies regarding negatives, temperature, or momentum on model performance. ... As shown in Table I, by estimating temperatures dynamically, GRACE+GLATE outperforms three models with static temperatures (τ=0.8, τ=0.2, or the same τ as GRACE). ... The results demonstrate that "τ=0.2", "τ=0.8", "w/ hard negative sampling", "w/ the same τ as GRACE", and GRACE+GLATE all converge rapidly, and our GRACE+GLATE model converges to the lowest loss value among them.

B. O2

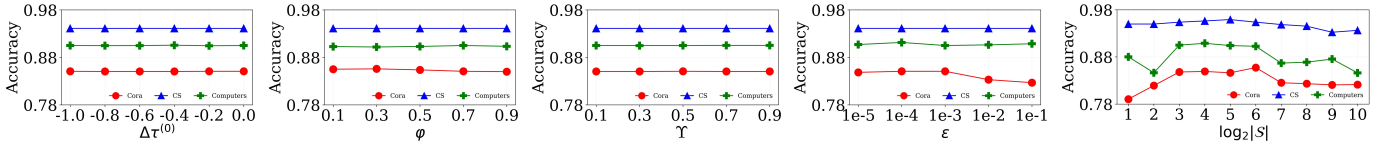
Comment: GLATE is only deployed with GRACE in the inductive node classification task. It would be beneficial to extend the application of GLATE to more baseline methods to demonstrate the effectiveness of the proposed approach.

Response: Thanks for your suggestion. We have conducted additional experiments by applying GLATE to other graph contrastive learning methods, including DGI, SUGRL, and NCLA, in the inductive node classification task. The results of these additional experiments have been included in Table II (i.e., Table VII in the revised manuscript) and show that GLATE consistently improves the performance of these baseline methods. Specifically, GLATE improves the test accuracy of DGI, SUGRL, and NCLA on PPI by 3.2, 1.8, and 1.8 percentage points, respectively. The updated results further validate the versatility and effectiveness of GLATE in the inductive node classification task.

The main revisions for the experimental analysis in the node classification task are as follows:

Raw texts: Table V shows the testing accuracy results on ~~the task of~~ transductive node classification. Firstly, we observe that several GCL models outperform the supervised GCN model on certain datasets, such as Coauthor-CS and Coauthor-Physics. ... Given that GRACE+GLATE consistently outperforms other models across all datasets, we select it as the representative model for GLATE in subsequent experiments.

The comparison results on ~~the~~ inductive node classification ~~task~~ are presented in Table VI. In addition, ~~GRACE+GLATE significantly outperforms all self-supervised baselines~~, highlighting the effectiveness of dynamic temperature estimation in inductive learning.



(a) Initial momentum (b) Momentum parameter (c) Similarity threshold (d) Temperature learning rate (e) Sampled node count

Fig. 2: Performance of GRACE+GLATE under different values of $\Delta\tau^{(0)}$, φ, Υ, ε, and $|\mathcal{S}|$ on Cora, CS, and Computers.

Revised texts: Table V shows the testing accuracy results on transductive node classification. Firstly, we observe that several GCL models outperform the supervised GCN model on certain datasets, such as CS and Physics. ... Given that GRACE+GLATE consistently outperforms other models on all datasets, we select it as the representative model for GLATE in subsequent experiments.

The comparison results on inductive node classification are presented in Table VI. ... In addition, GLATE improves the test accuracy of DGI, SUGRL, NCLA, and GRACE by 3.2, 1.8, 1.8, and 2.2 percentage points, respectively, highlighting the effectiveness of dynamic temperature estimation in inductive learning.

C. O3

Comment: More sensitivity analysis experiments are expected. There are several hyperparameters in the temperature definition, such as the initial momentum, the momentum parameter, the similarity threshold gamma and so on. However, the authors do not conduct enough parameter analysis in Section IV-4.

Response: As you suggest, we have expanded our sensitivity analysis to include additional experiments examining the effects of the initial momentum $\Delta\tau^{(0)}$, momentum parameter φ, similarity threshold Υ, temperature learning rate ε, and sampled node count $|\mathcal{S}|$ on the performance of GLATE. The experimental results on three different types of graph datasets (Cora: citation network, CS: co-authorship network, and Computers: co-purchase network) are shown in Fig. 2 (i.e., Fig. 4 in the revised manuscript). It verifies that the performance of GRACE+GLATE is robust to the changes in $\Delta\tau^{(0)}$, φ, and Υ, but sensitive to ε and $|\mathcal{S}|$. When $\varepsilon\in[10^{-4}, 10^{-3}]$ and $|\mathcal{S}|\in[8, 64]$, GLATE can achieve good performance. This general parameter configuration can achieve satisfactory performance when training GLATE on a new graph dataset without excessive parameter tuning time.

The main revisions for the experimental analysis of parameter sensitivity are as follows:

Raw texts: We conduct sensitivity analysis to assess the impact of ~~three crucial hyperparameters, namely~~ edge removing ratio p_r, attribute masking ratio p_m, and initial temperature $\tau^{(0)}$, on the performance of GRACE+GLATE. ... Furthermore, we observe that the setting of $\tau^{(0)}$=0.8 yields the highest accuracy result, reaching 84.5% on the test set. Given this finding, we recommend setting $\tau^{(0)}$=0.8 ~~on Cora~~. These conclusions are also applicable to the other datasets.

Revised texts: We conduct sensitivity analysis to assess the impact of edge removing ratio p_r, attribute masking ratio p_m, and initial temperature $\tau^{(0)}$, on the performance of GRACE+GLATE. ... This conclusion is also applicable to the other datasets. Furthermore, we observe that the setting of $\tau^{(0)}$=0.8 yields the highest accuracy result, reaching 84.5% on the test set. Given this finding, we recommend setting $\tau^{(0)}$=0.8. We also analyze the sensitivity of the initial momentum $\Delta\tau^{(0)}$, momentum parameter φ, similarity threshold Υ, temperature learning rate ε, and sampled node count $|\mathcal{S}|$ on Cora (citation network), CS (co-authorship network), and Computers (co-purchase network). Based on the results in Fig. 4, we can see that GRACE+GLATE is robust to the changes of $\Delta\tau^{(0)}$, φ, and Υ. Also, the results in Figs. 4 (d) and (e) reveal that $\varepsilon\in\left[10^{-4}, 10^{-3}\right]$ and $|\mathcal{S}|\in[8, 64]$ enable our method to achieve high test accuracy. This general parameter configuration can achieve satisfactory performance when training GLATE on a new graph dataset without excessive parameter tuning time.

D. O4

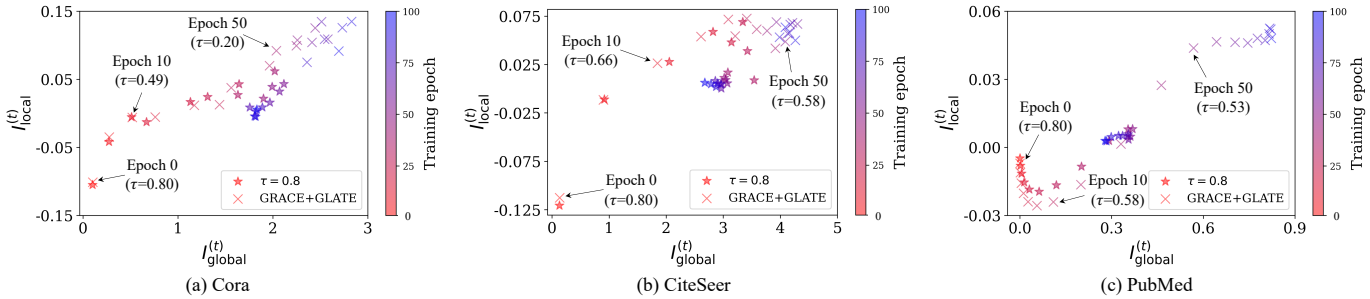
Comment: There are some minor issues in this paper. For example, the figures in Fig. 4 are not clear enough. The authors should ensure a clear separation between the points and text in each figure to avoid overlap. Furthermore, a clearer representation of how 'local' and 'global' change with increasing training epochs should be provided.

Response: Thanks for your kind reminder. We have revised the scatter plot (shown in Fig. 3, i.e., Fig. 5 in the revised manuscript) in the section of "Analysis of embedding quality" to ensure a clearer separation between points and text, thereby avoiding any overlap. Also, to better illustrate how "local" and "global" change with increasing training epochs, we use different colors to distinguish different training epochs, with a gradient color from red to blue representing training from epoch 0 to epoch 100. Furthermore, to present more clearly, we have reviewed the whole paper and adjusted other figure presentations in the paper. For example, we adjust the text in Fig. 4 (i.e., Fig. 10 in the revised manuscript) so that the scatter symbols and the text for the SIL score in each subfigure do not overlap and add subfigure number from (a) to (j) in Fig. 4.

IV. Response to Reviewer #3

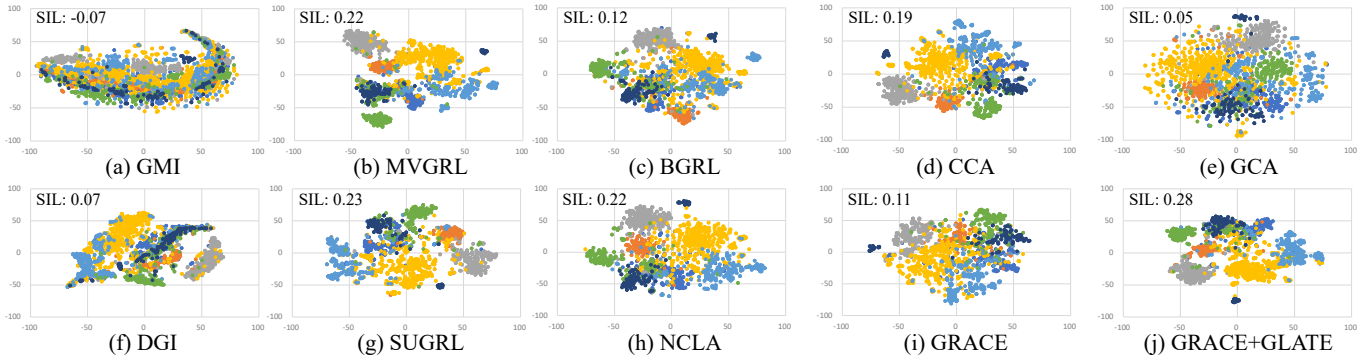
A. O1

Comment: Thanks for your comment. The innovative approach of dynamically estimating temperatures within contrastive losses, while theoretically sound, may pose challenges



(a) Cora (b) CiteSeer (c) PubMed

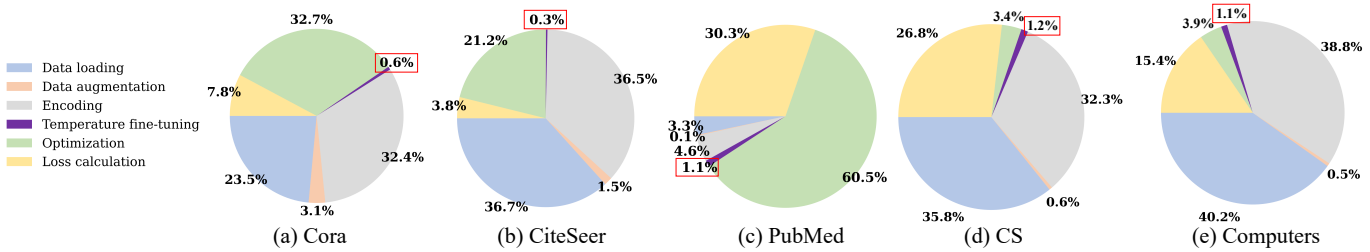
Fig. 3: Evolution trend of $I_{\text{global}}^{(t)}$ and $I_{\text{local}}^{(t)}$ during training. Red and blue represent early and late training epochs, respectively. The dots closer to the top right corner correspond to more reasonable CNEs.



(a) GMI (b) MVGRL (c) BGRL (d) CCA (e) GCA

(f) DGI (g) SUGRL (h) NCLA (i) GRACE (j) GRACE+GLATE

Fig. 4: Visualization of CNEs using t-SNE on Cora. Different colors represent different classes.



(a) Cora (b) CiteSeer (c) PubMed (d) CS (e) Computers

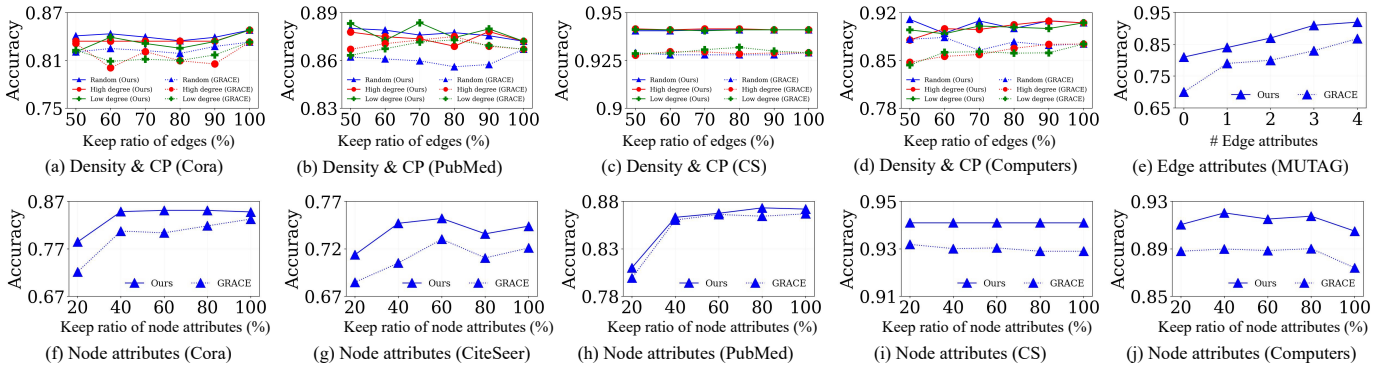
Fig. 5: Time ratio of different stages in the overall training of GRACE+GLATE. The temperature fine-tuning time ratio is marked with a red box. Different from the entire graph training on Cora, CiteSeer, CS, and Computers, GRACE+GLATE uses batch training on PubMed, so the loss calculation and optimization are relatively large.

for practical implementation. This complexity stems from the need to fine-tune the dynamic temperature parameters, which requires a deep understanding of the model's workings and the specific characteristics of the dataset being used. Such fine-tuning can be time-consuming and may necessitate extensive experimental trials to identify optimal settings.

Response: We have expanded the efficiency analysis of GLATE to provide a more comprehensive verification of the efficiency of temperature fine-tuning. Specifically: **(i)** We plot the time ratio of different stages in the overall training of GRACE+GLATE in Fig. 5. The results show that the time consumption of temperature fine-tuning is small in the overall training of GRACE+GLATE, not exceeding 1.2% on each dataset. **(ii)** Moreover, we have conducted comprehensive sensitivity analysis experiments on Cora (citation network), CS (co-authorship network), and Computers (co-purchase network). The results in Fig. 2 (i.e., Fig. 4 in the revised manuscript) show that GRACE+GLATE is robust to the changes of most hyperparameters, e.g., initial momentum and momentum parameter. Although GRACE+GLATE is sensitive to the temperature learning rate ε and sampled node count $|\mathcal{S}|$, the setting of $\varepsilon \in [10^{-4}, 10^{-3}]$ and $|\mathcal{S}| \in [8, 64]$ enables our method to achieve good results. This general parameter configuration can achieve satisfactory performance when training GLATE on a new graph dataset without excessive parameter tuning time.

The main revisions in the experimental analysis of model efficiency are as follows:

> **Raw texts**: By enabling the GCL model to learn high-quality CNEs in the training, GLATE achieves fast model convergence and thus minimizes the number of required training epochs. Based on the above analysis, we conclude that GRACE+GLATE has higher computational efficiency compared with the other GCL models.



(a) Density & CP (Cora) (b) Density & CP (PubMed) (c) Density & CP (CS) (d) Density & CP (Computers) (e) Edge attributes (MUTAG)

(f) Node attributes (Cora) (g) Node attributes (CiteSeer) (h) Node attributes (PubMed) (i) Node attributes (CS) (j) Node attributes (Computers)

Fig. 6: Performance of GRACE+GLATE under different graph structures and sizes. "CP" denotes connectivity pattern. "Keep ratio" refers to retaining a certain proportion of the edges or attributes from the original graph.

Revised texts: By enabling the GCL model to learn high-quality CNEs in the training, GLATE achieves fast model convergence and thus minimizes the count of required training epochs. Based on the above analysis, we conclude that GRACE+GLATE has higher computational efficiency compared with the other GCL models. In Fig. 7, we plot the time ratio of different stages during the training of GRACE+GLATE. We observe that the time consumption of temperature fine-tuning is small, not exceeding 1.2% on each dataset. Since GLATE updates $\tau^{(t)}$ based on sampled nodes at intervals of a few epochs during training, it does not impose a significant time burden on the original GCL model training.

B. O2

Comment: The paper's limited discussion on how different graph structures and sizes affect the model's performance leaves a gap in understanding its applicability across diverse real-world scenarios. Graphs can vary widely in terms of density, connectivity patterns, and node and edge attributes, each of which could significantly impact the effectiveness of the proposed method. A more comprehensive analysis, including a variety of graph types, would provide clearer insights into the model's versatility and limitations.

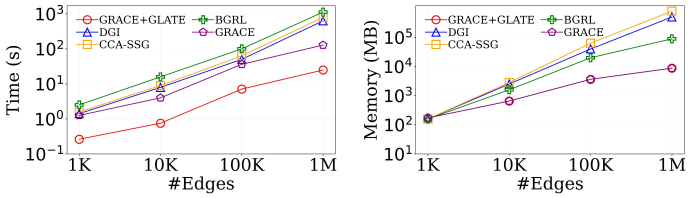
Response: Thanks for your suggestion. We have conducted additional experiments on a broader range of graph types with varying density, connectivity patterns, and node and edge attributes. These experiments include **(i)** varying graph density by removing different ratios of edges; **(ii)** varying graph connectivity pattern by setting different edge deletion priorities on different datasets; **(iii)** varying node attributes by masking different ratio of node attributes; **(iv)** varying node attributes by masking a different number of edge attributes. The results of these experiments are shown in Fig. 6 (i.e., Fig. 6 in the revised manuscript). The results reveal that compared to the base model GRACE, GRACE+GLATE exhibits better overall performance when changes occur in the graph structure and size or attribute information.

We have added a new section to our experiment, titled "Performance on Graphs with Different Structures, Sizes, and Attributes". The added content is as follows:

Revised texts: *Performance on graphs with different structures, sizes, and attributes:* Next, we analyze GRACE+GLATE's performance across different graph structures, graph sizes, and node (or edge) attributes, validated through density, connectivity patterns, node attributes, and edge attributes. Firstly, we validate density and connectivity patterns on Cora, PubMed, CS, and Computers, each with different connectivity patterns. We perform edge removing operations on each dataset, categorized into three types: random removing ("Random"), preferential removing of edges corresponding to high-degree nodes ("High degree"), and preferential removing of edges corresponding to low-degree nodes ("Low degree"). These operations result in three different connectivity patterns for each dataset. From Figs 6(a-d), we observe that compared to the base model GRACE, GRACE+GLATE exhibits better robustness on these datasets. It indicates that the temperature fine-tuning technique in GLATE enables the GCL model to learn good node embeddings even when the graph structure information is incomplete.

Secondly, since the MUTAG dataset contains edge attributes (i.e., four chemical bond types: aromatic, single, double, and triple bonds), we randomly mask edge attributes on it and observe GRACE+GLATE's performance as the scale of edge attributes changes. From Fig. 6(e), we see that complete edge attribute information allows the model to achieve high accuracy, and at the same time, GLATE remains more robust than GRACE when faced with reduced edge attributes.

Thirdly, we randomly mask node attributes on five datasets and observe GRACE+GLATE's performance as the scale of node attributes changes. From Fig. 6(f-j), we observe that both models exhibit a trend of overall performance degradation in citation networks as node attributes are missing; while in co-authorship networks and co-purchase networks, both models exhibit a trend of stable or improving performance as node attributes are missing. Moreover, when attribute information is missing, the performance of GRACE+GLATE is better than that of GRACE.



(a) Training time cost (b) GPU memory peak

Fig. 7: Results of the scalability experiments.

TABLE III: Comparison results on running time (seconds).

Model	Cora	CiteSeer	PubMed	Model	Cora	CiteSeer	PubMed
GMI	100.5	24.8	>500.0	DGI	10.5	*3.9*	158.2
MVGRL	21.6	13.5	>500.0	SUGRL	17.2	17.8	>500.0
BGRL	6.3	6.4	149.7	NCLA	>500.0	>500.0	>500.0
CCA-SSG	12.7	21.3	*89.9*	GRACE	*4.2*	4.5	349.9
GCA	11.4	4.2	315.4	GRACE+GLATE	**3.5**	**1.8**	**28.6**

C. O3

Comment: Given the computational complexity associated with dynamic temperature adjustments and the evaluation of embeddings, there could be scalability issues when applying the method to very large graph datasets. The paper does not fully address these potential challenges, particularly in contexts where computational resources are limited or where real-time processing is required. Scalability is a critical factor for many applications, and without a detailed examination of this aspect, it's difficult to assess the feasibility of deploying the proposed method in large-scale graph analytics tasks.

Response: Thanks for your comment. The findings presented in Fig. 5 (i.e., Fig. 7 in the revised manuscript) illustrate that the fine-tuning of temperature incurs minimal additional training time, suggesting that GLATE does not significantly increase the training load of the original GCL model. Furthermore, the application of a dynamic temperature enables the acquisition of high-quality embeddings more rapidly than a static approach, allowing GRACE+GLATE to achieve its objectives with fewer training epochs compared to GRACE alone, thereby shortening the overall training time, as demonstrated in Table III (i.e., Table XII in the revised manuscript). In addition, to examine the scalability of GLATE, we use the citation network of ogbn-arxiv to generate a group of graphs in different scales by sampling edges from it. The comparison of training time and peak GPU memory usage for different models is depicted in Fig. 7 (i.e., Fig. 8 in the revised manuscript). These findings reveal that the training time for GRACE+GLATE increases at a nearly linear rate as graph size expands, enhancing the time efficiency compared to GRACE. Moreover, both GRACE+GLATE and GRACE maintain a similar peak memory usage, surpassing other GCL baselines in terms of memory efficiency as the number of edges grows. The observations in Fig. 7 align with the insights from our complexity analysis in Section III-D, affirming that GRACE+GLATE is scalable to large-scale graphs.

We have added a new section to our experiment, titled "Scalability". The added content is as follows:

TABLE IV: Node clustering performance.

Model	Cora			CiteSeer			PubMed		
	Accuracy	NMI	ARI	Accuracy	NMI	ARI	Accuracy	NMI	ARI
Spectral clustering	36.7	12.6	3.1	23.8	5.5	1.0	52.8	9.7	6.2
k-means	49.2	32.1	22.9	54.0	30.5	27.8	59.5	*31.5*	28.1
GAE	59.6	42.9	34.7	40.8	17.6	12.4	*67.2*	27.7	27.9
VGAE	50.2	32.9	25.4	46.7	26.0	20.5	63.0	22.9	21.3
DGI	55.4	41.1	32.7	51.4	31.5	32.6	58.9	27.7	*31.5*
GRACE	*65.3*	*47.3*	*38.8*	*67.1*	*38.9*	*37.4*	56.2	20.2	19.9
GRACE+GLATE	**70.8**	**52.4**	**45.4**	**70.2**	**41.9**	**41.8**	**67.9**	**31.9**	**32.9**

TABLE V: Link prediction performance.

Model	UCI		Last.fm		Kuaishou	
	Recall@50	MRR	Recall@50	MRR	Recall@50	MRR
LINE	10.86	2.49	8.57	1.12	2.23	0.31
DGI	*25.49*	*5.74*	5.81	1.00	*5.11*	0.60
GraphCL	19.72	3.10	9.72	*1.51*	4.85	0.60
AD-GCL	18.19	3.23	8.22	1.11	1.84	0.24
GraphMAE	1.70	0.52	3.07	0.43	2.06	0.31
GRACE	16.69	2.91	*10.12*	1.45	4.68	*0.61*
GRACE+GLATE	**27.22**	**6.40**	**11.88**	**1.65**	**5.34**	**0.66**

Revised texts: *Scalability:* To examine the scalability of GLATE, we use the citation network of ogbn-arxiv to generate a group of graphs in different scales by sampling edges from it. The training time cost and GPU memory peak of different models are plotted in Figs. 8(a) and (b), respectively. We can see that with the increase in graph size, the time cost of GRACE+GLATE grows at a near-linear speed. Also, GRACE+GLATE and GRACE have almost the same memory peak, outperforming other GCL baselines in memory efficiency as edge count increases. Therefore, we conclude that GRACE+GLATE is scalable to large-scale graphs.

D. O4

Comment: The paper did not sufficiently explore the model's generalization capabilities across different graph-related tasks, such as link prediction, graph classification, or node clustering. Understanding how the proposed method performs in a variety of tasks is crucial for assessing its utility in real-world applications.

Response: Thanks for your suggestion. We have extended our evaluation to include additional graph-related tasks: in addition to the experiments on node classification and graph classification in the submitted manuscript, we have applied GLATE to node clustering and link prediction. In node clustering, we evaluated different methods on Cora, CiteSeer, and PubMed using the metrics of accuracy, normalized mutual information (NMI), and adjusted rand index (ARI). In link prediction, we evaluated different methods on UCI, Last.fm, and Kuaishou using the metrics of Recall@50 and mean reciprocal rank (MRR). The results of node clustering in Table IV (i.e., Table IX in the revised manuscript) show that GRACE+GLATE surpasses the base model GRACE by 6.7%, 6.6%, and 8.0% averagely in terms of accuracy, NMI, and ARI, respectively. The results of link prediction in Table V (i.e., Table X in the revised manuscript) show that GRACE+GLATE outperforms its base model GRACE by 4.3% and 1.2% averagely in terms of Recall@50 and MRR, respectively. The expanded

evaluation highlights the potential of our proposed method for a variety of graph-related tasks in real-world applications.

We have added two new sections to our experiment, titled "Performance on node clustering" and "Performance on link prediction". The added content is as follows:

Revised texts: *Performance on node clustering:* We further evaluate the performance of GLATE on node clustering using the metrics of accuracy, normalized mutual information (NMI), and adjusted rand index (ARI). Table IX shows that GRACE+GLATE surpasses the base model GRACE by 6.7%, 6.6%, and 8.0% averagely in terms of accuracy, NMI, and ARI, respectively, highlighting the benefits of dynamic temperature estimation to GCL in the node clustering task.

Revised texts: *Performance on link prediction:* We also evaluate the performance of GLATE on link prediction using the metrics of Recall@50 and mean reciprocal rank (MRR). Based on the results presented in Table X, we observe that GRACE+GLATE achieves superior performance in the link prediction task on all datasets. Specifically, GRACE+GLATE outperforms its base model GRACE by 4.3% and 1.2% averagely in terms of Recall@50 and MRR, respectively. These results demonstrate the effectiveness of incorporating dynamic temperature estimation into the GRACE model for link prediction.

Additionally, we have supplemented the dataset description and baseline description with relevant information on the datasets and baselines used in node clustering and link prediction.

Raw texts: For the node classification task, we use 10 commonly used benchmarks in the experiments, ... In a co-authorship network, each node represents an author, each edge represents the co-authorship between authors, and node features represent paper keywords for each author's papers. In a co-purchase network, each node represents a product whose features are bag-of-words from product reviews or product descriptions, and each edge indicates that two products are frequently bought together. ... ~~Additionally, we present the statistics of the above 10 datasets in Table III.~~

For the graph classification task, we use three datasets from TUDataset [50], including NCI1, MUTAG, and PROTEINS. NCI1 or MUTAG is grouped by small molecules, while PROTEINS is grouped by macromolecules. ~~In NCI1 or MUTAG, a small molecule corresponds to a graph where nodes and edges are atoms and chemical bonds, respectively. In PROTEINS, a macromolecule corresponds to a graph where nodes are secondary structure elements. If two nodes are connected, they are neighbors along the amino acid sequence or one of the three nearest neighbors in space.~~ Furthermore, we present the statistics of the three datasets in Table IV.

TABLE VI: Statistics of the datasets for link prediction.

Dataset	# Node	# Edge	Link
UCI	1,677	56,617	http://konect.unikoblenz.de/networks/
Last.fm	127,786	720,537	http://www.last.fm/
Kuaishou	138,812	1,779,639	http://www.kuaishou.com/

Revised texts: **(i)** For the node classification task, we use 10 commonly used benchmarks in the experiments, ... In a co-authorship network, each node represents an author, each edge represents the co-authorship between authors, and node attributes represent paper keywords for each author's papers. In a co-purchase network, each node represents a product whose attributes are bag-of-words from product reviews or product descriptions, and each edge indicates that two products are frequently bought together. ... **(ii)** For the graph classification task, we use three datasets from TUDataset [62], including NCI1, MUTAG, and PROTEINS. NCI1 or MUTAG is grouped by small molecules, while PROTEINS is grouped by macromolecules. **(iii)** For the node clustering task, we use the datasets of Cora and CiteSeer as benchmarks. **(iv)** For the link prediction task, we use three datasets: UCI (message communications among University of California, Irvine students in an online community), Last.fm (<user, artist, song> tuples from the Last.fm API, representing the listening habits of nearly 1,000 users), and Kuaishou (interactions between 6,840 users and 131,972 videos from the Kuaishou online video-watching platform). We present the statistics of the above datasets in Tables III, IV, and V.

We also added the statistics of the used datasets for link prediction in Table VI (i.e., Table V in the revised manuscript).

Accordingly, we have revised the descriptions of the experiments in the abstract, contributions and conclusion:

Raw texts: Finally, the extensive experiments on 13 benchmark datasets demonstrate that GLATE consistently outperforms the state-of-the-art graph contrastive learning models on either classification accuracy or training efficiency.

Revised texts: Finally, the extensive experiments on 16 benchmark datasets demonstrate that GLATE consistently outperforms the state-of-the-art graph contrastive learning models in terms of both model performance and training efficiency.

Raw texts: Through the extensive experiments on 13 benchmark datasets, we demonstrate that GLATE significantly improves the performance of five representative GCL base models on downstream tasks, including node classification and graph classification (Section IV).

Revised texts: Through the extensive experiments on 16 benchmark datasets, we demonstrate that GLATE significantly improves the performance of five representative GCL base models on downstream tasks, including node classification, graph classification, node clustering, and link prediction (Section IV).

Raw texts: Our experiments on two important graph-related tasks, node classification and graph classification, demonstrate the significant improvements of GLATE in classification accuracy and training efficiency compared with the state-of-the-art graph contrastive learning models. ~~Furthermore, we validate the high-quality contrastive node embeddings learned by GLATE using the indicators of global separation and local separation.~~ In future work, we plan to extend the GLATE method to distributed environments, exploring strategies to leverage parallel processing and efficient communication protocols.

Revised texts: Our experiments on four graph-related tasks demonstrate the significant improvements of GLATE in model performance and training efficiency compared with the state-of-the-art graph contrastive learning models. In future work, we plan to extend the GLATE method to distributed environments, exploring strategies to leverage parallel processing and efficient communication protocols.

E. O5

Comment: While the method focuses on improving the quality of contrastive node embeddings, there is limited discussion on the interpretability of these embeddings. For practical applications, especially in domains like bioinformatics or social network analysis, the ability to interpret and understand the embeddings can be as important as their performance.

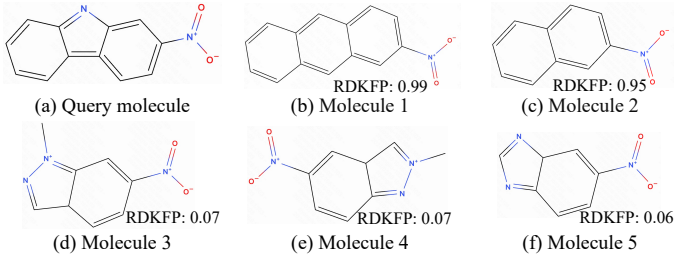
Response: Thanks for your comment. We have added a new section (i.e., Section IV-B-11) to the revised manuscript that discusses the interpretability of the embeddings generated by our method. Specifically, for a query molecule in MUTAG, we identified its two closest molecules and three farthest molecules based on embedding distance. The results in Fig. 8 (i.e., Fig. 9 in the revised manuscript) show that the embedding distance is closely related to the RDKit fingerprint similarity between chemical compounds, which indicates that the embeddings learned by GRACE+GLATE match the basic chemical knowledge of molecules.

We have added a new section to our experiment, titled "Case study". The added content is as follows:

Revised texts: *Case study:* To interpret the learned embeddings, we identify the two closest molecules and three farthest molecules in terms of embedding distance from a query molecule in MUTAG. From Fig. 9, we find that the distances between pairs of embeddings closely correspond to the computed similarity between chemical compounds based on their RDKit fingerprints (RDKFP). For example, since the query molecule and molecule 1 (or 2) both have at least two benzene rings, they have high similarity both in terms of embedding similarity and RDKFP similarity. It implies that the learned embeddings match the basic chemical knowledge of molecules.

F. O6

Comment: This work appears more aligned with AI/ML research themes, as indicated by its heavy citation of sources from AI/ML conferences and journals, which may not perfectly fit the traditional focus areas of ICDE.



(a) Query molecule (b) Molecule 1 (c) Molecule 2
(d) Molecule 3 (e) Molecule 4 (f) Molecule 5

Fig. 8: Query molecule and its most similar (molecules 1 and 2) / least similar (molecules 3, 4, and 5) molecules based on the embeddings learned by GRACE+GLATE on MUTAG.

Response: Thanks for your comment. In this paper, we focus on improving the training process of graph contrastive learning (GCL). It is important to note that GCL supports downstream tasks, which are integral to the field of data mining. For example, node classification is important for social/citation network analysis; graph classification is important for molecule property prediction; link prediction is important for recommender systems. In the revised manuscript, we apply GLATE to four different downstream tasks, including node classification, graph classification, node clustering, and link prediction. Our proposed method GLATE not only improves the performance of the base model but also accelerates its training process, which makes GCL models easily scalable to large-scale graph data.

Furthermore, we have investigated the works closely related to ours that were published in SIGMOD, VLDB, and ICDE, and added 16 new references into the revised manuscript. These papers cover self-supervised learning ([1], [2], [3], [21], and [22]), node classification or link prediction on temporal graphs ([29], [31], [32], [38] and [39]), citation network analysis ([30], [33], and [36]), molecule property prediction ([34] and [35]), and link prediction on knowledge graphs ([37]).

The main revisions for the citations in the revised manuscript are as follows:

Raw texts: Self-supervised learning provides a promising learning paradigm without relying on high-cost label information for many research fields such as computer vision [1]–[3], natural language processing [4]–[7], speech recognition [8]–[10], and recommender systems [11]–[13]. Contrastive-based methods have a prominent place among the landscape of self-supervised learning methods [14]–[17]. Contrastive learning leverages the inherent structure and relationships within unlabeled data to train encoder networks [18]–[20]. ... On benchmark datasets, the state-of-the-art GCL models have demonstrated competitive performance against supervised learning models, e.g., graph convolutional network (GCN) [26], in various graph-related tasks such as node classification [23], [27] and graph classification [28], [29]

Revised texts: Self-supervised learning [1-3] provides a promising learning paradigm without relying on high-cost label information for many research fields such as computer vision [4]–[6], natural language processing [7]–[10], speech recognition [11]–[13], and recommender systems [14]–[16]. Contrastive-based methods have a prominent place among the landscape of self-supervised learning methods [17]–[20]. Contrastive learning leverages the inherent structure and relationships within unlabeled data to train encoder networks [21], [22]. ... On benchmark datasets, the state-of-the-art GCL models have demonstrated competitive performance against supervised learning models, e.g., graph convolutional network (GCN) [28], in various graph-related tasks such as node classification [29-33], graph classification [34], [35], and link prediction [36-39].

The specific references that have been newly added into the revised manuscript can be found on page 2.