Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36443

GeoGRACE: Geometry-Inspired Low-Rank Parameter Perturbation for Efficient Graph Contrastive Learning

Xueyi Tan

School of Mathematical Sciences, Peking University, Beijing, China
2300010816@stu.pku.edu.cn

Abstract. In molecular property prediction, unsupervised graph contrastive learning (GCL) is pivotal when obtaining labeled data is costly. However, conventional augmentation-free approaches based on parameter perturbation mainly rely on full-rank isotropic noise, which can introduce uncontrolled perturbations and impair semantic consistency of the molecular representations. To mitigate semantic degradation and improve representation consistency, this paper therefore introduces GeoGRACE, a geometry-inspired low-dimensional parameter perturbation framework that generates contrastive views by controlling the amount of noise injected into the parameters in a low-dimensional random subspace using low-rank matrix factorizations. Experiments on five biochemical benchmark datasets show that this structured geometric prior offers improved or comparable downstream classification accuracy over the SimGRACE baseline, for example, improving accuracy from 78.35% to 79.18% on NCI1 dataset. The robustness analysis suggests that early stopping and layer-wise adaptive rank can improve performance stability. Moreover, introducing additional low-rank factorization operations does not necessarily introduce substantial computational overhead; the Layer-wise Adaptive Rank variant, for example, reduces training time by up to 18.5% with manageable memory cost. This work presents a computationally efficient and representation-stable pre-training paradigm that shows promise for biochemical graph representation learning.

Keywords: graph contrastive learning, parameter perturbation, low-rank factorization, unsupervised learning, biochemical molecules

1. Introduction

Graph neural networks (GNNs) have become the de facto approach to relational data modeling and molecular property prediction [1]. In the pharmaceutical industry, especially for disease-related tasks, molecular structures are naturally represented as graph data, where atoms are nodes and chemical bonds are edges. However, a real-world challenge persists: while there are vast repositories of unlabeled molecular graphs, acquiring labels for specific properties such as toxicity, efficacy or targeted disease curation, is costly and time-consuming [2]. Hence, there is a pressing need for unsupervised graph contrastive learning (GCL) approaches that enable AI systems to learn informative molecular representations from molecules prior to task-specific fine-tuning [3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.36443

Most of the work in the field of GCL focuses on the creation of good quality contrastive views using data augmentation. Conventional approaches, like GraphCL, often use high-level augmentations on graphs, including randomly dropping nodes or sampling sub-graphs [4]. Models such as JOAOv2 suggest unified bi-level optimization for the selection of augmentation that can automate this process and remove manual trial-and-error [5]. These generic augmentations, which can be successfully used in other domains such as computer vision, may be problematic for biochemical data where they can change chemical structures or lack consistency in representations. A way to reduce this semantic instability is to use expensive domain knowledge, like MoCL which substitutes valid substructures with bioisosteres [6]. Other recent techniques do not resort to architectural modifications but rather add Gaussian noise directly to the parameters of the encoder, like SimGRACE, or use asymmetric neighbor aggregation, like GraphACL [7, 8]. However, one problem remains: when using full rank, isotropic Gaussian noise, all directions of the parameters are perturbed indiscriminately, introducing redundant perturbations in less relevant parameter directions. This blind perturbation can lead to a lower capability of capturing the core functional groups in the delicate encoding space of molecules and result in lower semantic consistency of the generated views.

The motivation of this paper is to develop an efficient and stable GCL framework to tackle the problems in a systematic manner. To balance between the two, this paper introduces a new framework named GeoGRACE, which uses structured low-rank noise instead of isotropic noise for parameter perturbations. To demonstrate the performance of the proposed method, experiments are conducted on five biochemical benchmark datasets: MUTAG, PROTEINS, DD, NCI1 and AIDS. In particular, this study compares GeoGRACE and its variants against the full-rank isotropic Gaussian noise baseline (SimGRACE) from three main aspects: (1) whether low-dimensional geometric spaces, in which perturbations of parameters are restricted, yield an improvement over the full-rank isotropic Gaussian noise baseline; (2) how sensitive this framework that is based on geometry can be to its fundamental hyperparameters, such as the scale of the perturbations; and (3) if structured low-dimensional perturbations can maintain computational efficiency (training time and memory) without compromising the quality of the results. These methods are detailed in Section 2, which introduces the data sources and explains the proposed model and algorithm. Section 3 includes the experimental analysis and Section 4 summarizes the findings and sets out limitations and future work.

2. GeoGRACE framework

The augmentation-free GCL paradigm is the basis for the proposed GeoGRACE framework. The overall methodology is shown in Figure 1. The idea is to give more semantically stable contrastive views of the original molecular graph to reduce destructive transformations of the input data. In particular, GeoGRACE produces two correlated molecular representations with a dual-encoder. It proposes structured and low-rank perturbations on the parameters of an encoder in low-dimensional parameter subspaces instead of perturbing the parameters with full rank Gaussian noise.

2.1. Dual-encoder architecture and low-rank perturbation algorithm

A Graph Isomorphism Network encoder parameterized by weights θ aggregates node-level embedding $h = f(\mathcal{G};\theta)$ given a graph $\mathcal{G}$ [9]. At the same time, an altered version of the encoder $f(\cdot;\theta')$ encodes $\mathcal{G}$ to get a correlated representation $h' = f(\mathcal{G};\theta')$. Both representations are projected into a non-linear latent space, as a result of a non-linear 2-layer multi-layer perceptron

DOI: 10.54254/2755-2721/2026.36443

projection head $g(\cdot)$, obtaining the final embeddings $z = g(h)$ and $z^{'} = g\left(h^{'}\right)$ for contrastive optimization. For every iteration of training, $2N$ representations of the graphs are produced from N minibatch graphs. This study uses the normalized temperature-scaled cross-entropy loss (NT-Xent) to maximize the mathematical agreement between the positive pair $\left(z_n, z_n^{'}\right)$ originating from the same graph and push all negative pairs, in the batch, apart [10]. The loss function for the n -th graph is written as:

$$l_n = -log\frac{e^{sim\left(z_n, z_n^{'}\right)/\tau}}{\sum_{n^{'}=1, n^{'}\neq n}^{N} e^{sim\left(z_n, z_n^{'}\right)/\tau}} \tag{1}$$

where $\text{sim}(u, v) = \frac{u^T v}{\|u\| \|v\|}$ is the cosine similarity function and τ is a temperature hyperparameter to control the sensitivity of the distribution.

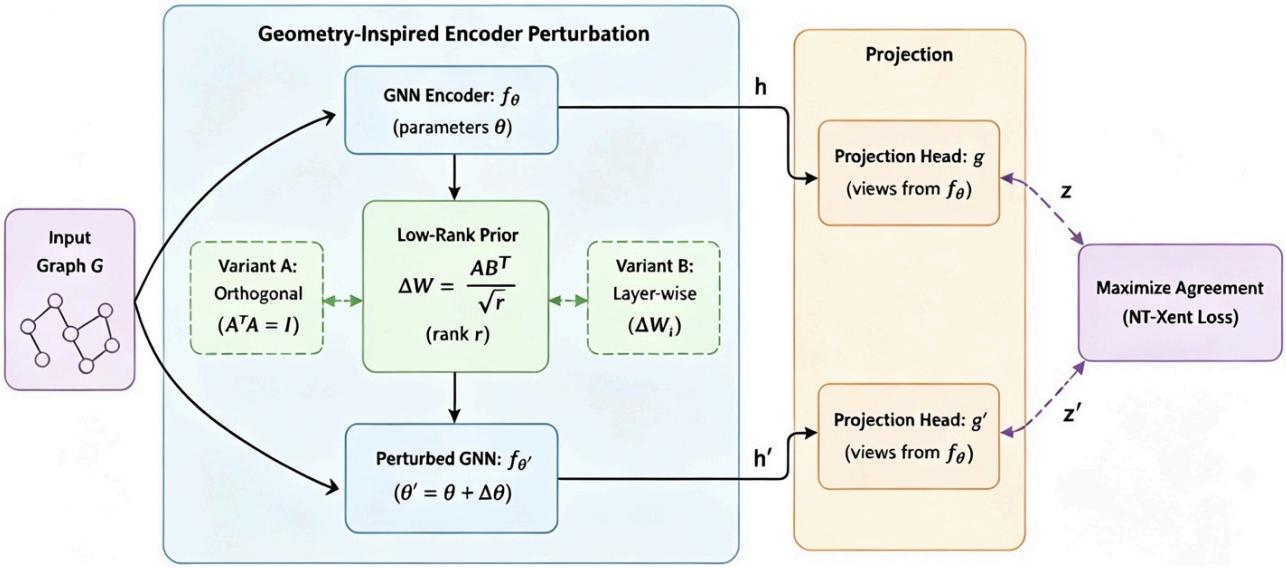


Figure 1. The overall architecture of the GeoGRACE framework

2.2. Geometry-inspired low-rank parameter perturbation

The key contribution of GeoGRACE is the perturbed encoder's generation logic $f\left(\cdot;\theta^{'}\right)$. SimGRACE baseline produces θ' using isotropic Gaussian noise $\theta' = \theta + \eta \cdot \Delta\theta$, where $\Delta\theta \sim \mathcal{N}\left(0, \sigma^2\right)$ and η regulates the magnitude of the perturbation. This full-rank perturbation, however, can by itself impact the semantic structures in biochemical molecular graphs without considering the geometric structure. The Random Low-Rank Perturbation (RoRA) constraint can be used to overcome this. Typically, network parameters may lie in low-dimensional subspaces and so unconstrained perturbations can lead to perturbations in directions not important for learned representation [11]. Thus, for arbitrary $2D$ weight matrix $W \in \mathbb{R}^{d_{out} \times d_{in}}$ in the linear layers of the GNN encoder, this study restricts the perturbation noise to a low-dimensional subspace and randomly samples 2 low-rank factor matrices $A \in \mathbb{R}^{d_{out} \times r}$ and $B \in \mathbb{R}^{r \times d_{in}}$ generated from $\mathcal{N}(0,1)$. Rank r is a hyperparameter carefully chosen to be noticeably less than the dimensions of the original

DOI: 10.54254/2755-2721/2026.36443

matrix, mathematically speaking, $r \ll \min(d_{out}, d_{in})$. These two factors are then multiplied to produce the structured low rank noise:

$$\Delta W = \frac{1}{\sqrt{r}} AB^T \tag{2}$$

In this formulation, the variance scaling is done mathematically by dividing the matrix product by the square root of the rank ($\sqrt{r}$). This normalization provides a stable and consistent expected perturbation magnitude, and compares with the unconstrained baseline regardless of the chosen rank r. The degrees of freedom of the whole perturbation mechanism are constrained by using ΔW instead of the isotropic noise $\Delta\theta$. This preserves the mapping power of the GNN while enabling more stable representations of molecules, and still provides enough mathematical variation and view diversity to facilitate the contrastive learning process.

2.3. Two advanced variants of geometric constraints

The Orthogonally Initialized Low-Rank variant is designed to address the potential issue of redundant parameter shifts caused by randomly sampled, overlapping directional vectors. It uses orthonormal bases to initialize the factor matrices A and B, instead of the standard Gaussian distributions. This helps to ensure the potentially increasing directional diversity in a direction and to avoid unnecessary parameter changes. Additionally, the Layer-Wise Adaptive Rank variant exponentially reduces the pre-specified rank r with the forward pass as it gets deeper. It uses more geometric exploration in shallow feature-extraction layers while preserving more abstracted, high-level semantics in deeper layers, which may help to alleviate GNN over-smoothing [12].

3. Experimental setup and results

3.1. Datasets and experimental protocol

Experiments are performed on five well-known biochemical molecular graph datasets from the benchmark TUDataset, including MUTAG, PROTEINS, DD, NCI1, and AIDS [13]. Table 1 gives a comprehensive summary of these benchmark datasets.

Table 1. Biochemical datasets MUTAG, PROTEINS, DD, NCI1 and AIDS statistics

Statistics	MUTAG	PROTEINS	DD	NCI1	AIDS
Graphs	188	1113	1178	4110	2000
Classes	2	2	2	2	2
Avg. Nodes	17.93	39.06	284.32	29.87	15.69
Avg. Edges	19.79	72.82	715.66	32.30	16.20

These datasets present diverse biochemical tasks, classifying mutagenicity (MUTAG), protein structures (PROTEINS, DD), and medical activities (NCI1, AIDS).

The performance of the representations is evaluated in terms of classification performance on data from downstream tasks, evaluating the mean and standard deviation of the test accuracy with 5 random seeds on 10-fold stratified cross-validation with statistical significance assessed via Welch's

DOI: 10.54254/2755-2721/2026.36443

t-test. All models were run on Apple Silicon (MPS) GPUs, training for 20 or 50 epochs with a fixed learning rate of 0.01 and a fixed batch size of 128 and using a simple grid search of r and η.

3.2. Representation quality: downstream classification performance

The representation quality is determined by the performance of the downstream classification tasks. The maximum mean test accuracy and its standard deviation were calculated from a hyperparameter search including different perturbation scales ($\eta \in \{0.1,1,10\}$), ranks ($r \in \{4,8,16\}$). The summary of these best mean results, comparing GeoGRACE against the SimGRACE are shown in Table 2.

Table 2. Performance comparison on biochemical datasets (best mean acc % ± std)

Dataset	SimGRACE	RoRA	Orthogonal	Layer-wise
MUTAG	87.87 ± 1.98	89.17 ± 1.22	88.43 ± 1.39	89.56 ± 2.57
PROTEINS	74.76 ± 0.78	75.47 ± 0.59	75.17 ± 0.53	74.97 ± 0.38
DD	75.94 ± 1.03	76.72 ± 0.52	76.40 ± 0.55	76.41 ± 0.66
NCI1	78.35 ± 0.94	79.18 ± 0.50	78.76 ± 0.74	78.92 ± 0.45
AIDS	99.56 ± 0.05	99.64 ± 0.07	99.61 ± 0.09	99.63 ± 0.07

Empirical results indicate that low-dimensional geometric perturbations of parameters yield, overall, improved or competitive performance on the various datasets. On PROTEINS, DD, AIDS and NCI1, the accuracy of the RoRA perturbation was better. Interestingly, in the large-scale NCI1, the geometrically constrained approach achieved a better score than the baseline, improving from 78.35% to 79.18%. Although the standard deviations overlap to some degree—which is common in contrastive learning because of its unconditional sensitivity to seed—the competitive mean accuracies suggest that RoRA can provide a balance between perturbation diversity and representation quality. On AIDS dataset, the improvement is marginally significant according to Welch's t-test ($p \approx 0.075$).

Notably, the highest accuracy score of 89.56% is obtained by the Layer-wise Adaptive Rank variant on MUTAG. This phenomenon suggests that the size of the perturbations applied across the layers could be dynamically changed, matching the topological shape of some of the molecular graphs, which would contribute to keeping representation diversity in deeper layers.

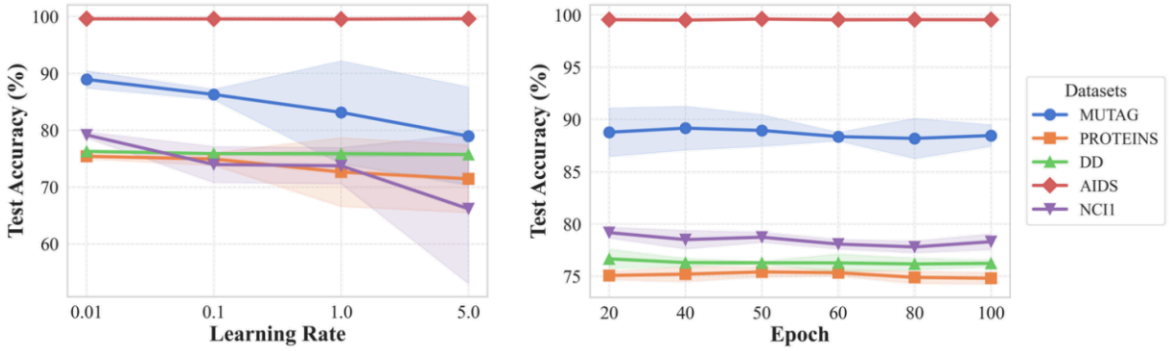
3.3. Robustness analysis: hyperparameter sensitivity

To assess the stability and robustness of RoRA, this study performed a sensitivity analysis of the hyperparameters. Figure 2 shows the impact of learning rate and training epochs on downstream performance, and Figure 3 shows the framework's sensitivity to the rank r and perturbation scale η.

As shown in Figure 2(a), the model seems to perform best at low rates (0.01) and high rates (1.0 or 5.0) result in a discernible loss of accuracy. This suggests that overly aggressive learning rates may push the encoder away from the semantic manifold of molecular graphs. Moreover, Figure 2(b) plots representation quality against a number of epochs trained. The augmentation-free GCL framework has the advantage that it can be early stopped as maximum performance is typically reached after 20 to 50 epochs. In experiments, no significant further improvement was observed

DOI: 10.54254/2755-2721/2026.36443

with longer training, suggesting that excessive training may reduce generalization under the generated perturbation distribution than the underlying topological structures.



(a) Impact of learning rate (b) Impact of training epochs

Figure 2. The effect of learning rate and training epochs on the RoRA framework

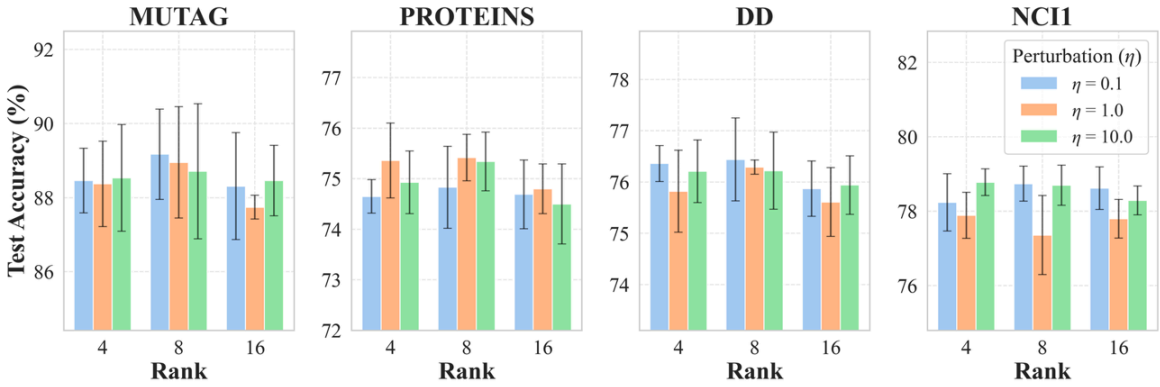


Figure 3. The effect of rank r and perturbation scale η on the RoRA framework

There is a clear trade off: the more aggressive the ranks (e.g. $r < 4$), the less space for the update, and the less views are diversified. In contrast, the structured noise can be seen to resemble the full rank isotropic baseline as r approaches 16. Regarding the perturbation scale, for different levels of η (0.1, 1.0, 10.0), the system is found to be robust. RoRA's accuracy remains stable even under extreme noise, whereas unconstrained baselines may become more sensitive to large perturbation scales. Across all datasets, the moderate ranks ($r = 4$ or $r = 8$) and conservative perturbation scales ($\eta = 0.1$ or 1.0) yield the highest accuracy. Lower ranks ($r = 4$) often perform best on smaller datasets like PROTEINS, while higher ranks ($r = 16$) become competitive on larger datasets like DD.

3.4. Computational efficiency: training time reduction and memory bounding

One often mentioned problem with adding algebraic operations to the pre-training pipeline is that it may add to the computational load. The average empirical training time and peak memory consumption with the 20 and 50 training epochs were calculated and shown in Table 3.

Contrary to the expectation that the more algebraic constraints the longer it takes to train, the proposed RoRA method reduced training time across all evaluated datasets. The observed hardware speedup likely stems from the lower computational cost of the perturbation matrices and

DOI: 10.54254/2755-2721/2026.36443

implementation-level optimizations. For instance, on the large-scale NCI1 dataset at Epoch 20, the Layer-wise variant reduced training time from 34.51s to 28.13s, achieving an 18.5% speedup.

Table 3. Computational efficiency comparison (averaged time in seconds / peak memory in MB)

Dataset	Epochs	SimGRACE	RoRA	Orthogonal	Layer-wise
MUTAG	20	1.95 / 639.2	1.78 / 639.5	2.15 / 641.3	1.85 / 642.2
	50	4.83 / 639.7	4.22 / 640.3	5.08 / 641.8	4.37 / 643.1
PROTEINS	20	8.80 / 729.4	8.40 / 731.2	10.12 / 733.4	7.96 / 733.9
	50	22.13 / 732.8	19.71 / 733.3	24.95 / 736.5	19.37 / 736.6
DD	20	30.26 / 1063.0	30.00 / 1122.5	31.66 / 1121.3	29.08 / 1085.5
	50	75.87 / 1068.9	73.69 / 1134.4	78.60 / 1132.1	71.37 / 1087.4
NCI1	20	34.51 / 1065.5	30.47 / 1064.0	34.84 / 1066.6	28.13 / 1068.2
	50	83.99 / 1076.0	80.19 / 1075.0	85.20 / 1079.6	77.87 / 1079.8
AIDS	20	16.43 / 813.1	15.45 / 813.2	16.67 / 815.3	14.51 / 816.0
	50	41.05 / 818.6	36.45 / 818.3	43.05 / 821.3	36.15 / 821.3

The exploratory variants also shed light on the effects of geometric constraints on computation efficiency. Overall, the Layer-wise Adaptive Rank variant was the fastest. The dimensions of the generated perturbation matrices decrease exponentially as the network gets deeper, and the calculation burden of deep layers during forward passes is also reduced directly (e.g. DD Epoch 50 training from 75.87s to 71.37s). On the other hand, Orthogonally Initialized Low-rank variant required slightly longer training times, such as 50 epochs on PROTEINS, which was increased to 24.95s (an acceptable trade-off, since a strict orthogonal basis initializations were required).

As for spatial complexity, memory use was kept to acceptable levels. For MUTAG, PROTEINS, NCI1 and AIDS, overheads were negligible. For DD, while average memory increased moderately (from 1063.0MB to 1122.5MB), careful hyperparameter selection (e.g., $r = 8$, $\eta = 1.0$) can reduce the overhead to 1088.2MB, demonstrating that memory cost is controllable and not a bottleneck.

4. Conclusion

This paper highlights a limitation of current augmentation-free (GCL) frameworks that rely on full-rank, isotropic Gaussian noise, which may lead to space warping and affect the semantics in biochemical molecular graphs. After introducing the GeoGRACE framework and its core RoRA algorithm, it is found that mathematically restricting parameter noise in a low-dimensional random subspace improves downstream performance while maintaining stable training behavior. The empirical evaluations provide three findings: (1) GeoGRACE generally improves the accuracy of downstream classification tasks, indicating that a low-rank geometric prior is useful for maintaining representation quality; (2) the framework benefits from early stopping, and the layer-wise adaptive rank mechanism may alleviate representation degradation; and (3) compressing the space of

DOI: 10.54254/2755-2721/2026.36443

parameter updates reduces training time while keeping memory overhead negligible on most datasets.

The results also provide a promising way for unsupervised graph representation learning. The work introduces an alternative technique that does not need costly augmentations specific to the domain and also does not just use semantically blind noise injections. A primary contribution of this work is the low computational cost pre-training paradigm that could serve as a lightweight perturbation module for researchers of computational chemistry and pharmaceutical AI.

However, there are some limitations in this current study. The choice of predefined rank r is currently based on grid search and empirical tuning, which could lead to extra tuning workload when using the model on totally new datasets. Moreover, the experiments were only verified in the framework of biochemical graphs. Thus, future research will focus on automated and data-driven approaches for estimating the rank, without relying on manual hyperparameter tuning. Moreover, investigating the transferability of this low rank encoder perturbation, inspired by geometry, to other complex and large-scale architectures like Large Language Models.

References

[1] He, Y., Zheng, Z., Soga, P., Zhu, Y., Dong, Y., & Li, J. (2024). Explaining graph neural networks with large language models: A counterfactual perspective on molecule graphs. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 7079-7096.

[2] Yang, S., & Deng, L. (2026). Self-supervised learning for molecular property prediction: Methods, multimodal insights, and benchmark comparisons. *Journal of Chemical Information and Modeling, 66*(15), 8705-8734.

[3] Zhang, X., Xu, Y., Jiang, C., Shen, L., & Liu, X. (2024). MoleMCL: A multi-level contrastive learning framework for molecular pre-training. *Bioinformatics, 40*(4), btae164.

[4] You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., & Shen, Y. (2020). Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems, 33*, 5812-5823.

[5] You, Y., Chen, T., Shen, Y., & Wang, Z. (2021). Graph contrastive learning automated. *Proceedings of the 38th International Conference on Machine Learning, 139*, 12121-12132.

[6] Sun, M., Xing, J., Wang, H., Chen, B., & Zhou, J. (2021). MoCL: Data-driven molecular fingerprint via knowledge-aware contrastive learning from molecular graph. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 3585-3594.

[7] Xia, J., Wu, L., Chen, J., Hu, B., & Li, S. Z. (2022). SimGRACE: A simple framework for graph contrastive learning without data augmentation. *Proceedings of the ACM Web Conference 2022*, 1070-1079.

[8] Xiao, T., Zhu, H., Chen, Z., & Wang, S. (2023). Simple and asymmetric graph contrastive learning without augmentations. *Advances in Neural Information Processing Systems, 36*, 16129-16152.

[9] Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks? arXiv preprint arXiv:1810.00826.

[10] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning, 119*, 1597-1607.

[11] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[12] Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., & Sun, X. (2020). Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(4), 3438-3445.

[13] Morris, C., Kriege, N. M., Bause, F., Kersting, K., Mutzel, P., & Neumann, M. (2020). TUDataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*.