Exploring Enhancing Opportunities of MolCLR: Molecular Contrastive Learning of Representations via Graph Neural Networks

Janindu Kulathunga
dept. of Computer Science and Engineering
University of Moratuwa, Sri Lanka
janindu.21@cse.mrt.ac,lk

***Abstract*—MolCLR has firmly established itself as a leading framework for self-supervised molecular representation learning, leveraging the power of graph neural networks in conjunction with contrastive learning. Its capacity to encode rich molecular features without relying on extensive labelled datasets has rendered it highly effective across a range of chemical and pharmaceutical applications. Nevertheless, despite its considerable success, there remain several avenues for improvement to further enhance its predictive accuracy, generalisation capability, and robustness in downstream tasks. In this work, we propose a series of incremental yet practical enhancements to the original MolCLR framework. These improvements encompass advanced optimisation strategies, such as the adoption of AdamW and sophisticated learning rate schedules, architectural modifications including the integration of graph transformers and substructure-based attention mechanisms, and chemically informed data augmentations that preserve molecular semantics while introducing meaningful diversity in the training set. By systematically evaluating these enhancements on established MoleculeNet benchmarks, we observe notable gains in representation robustness, task generalisation, and predictive performance. These findings highlight the potential of the proposed modifications to extend the utility of MolCLR in molecular property prediction, providing a more reliable and versatile framework for self-supervised learning in cheminformatics.**

***Index Terms*—MolCLR, Molecular representation learning, Self-supervised learning, Graph neural networks (GNNs), Contrastive learning, Graph transformers, Substructure attention, Data augmentation, Cheminformatics, Molecular property prediction, Optimisation strategies, MoleculeNet benchmarks**

I. Introduction

Self-supervised molecular representation learning has increasingly become a cornerstone in modern drug discovery and computational chemistry, primarily due to the limited availability of high-quality labelled datasets and the high cost associated with experimental data generation. Accurate molecular embeddings enable models to generalise across a wide range of chemical tasks, from predicting physicochemical properties to assessing bioactivity and toxicity. MolCLR [2] represents a significant advancement in this domain, employing contrastive learning in combination with graph neural networks (GNNs) to learn rich, transferable molecular representations from unlabelled molecular graphs. These embeddings can then be fine-tuned for diverse downstream tasks, offering a practical solution to the challenge of data scarcity. Despite its effectiveness, the original MolCLR framework leaves several avenues unexplored, particularly in the areas of optimisation strategy, architectural design, and chemically informed data augmentation, which can further enhance the quality and generalisability of the learned representations. This paper presents a systematic approach to improving MolCLR by introducing advanced optimisation techniques such as AdamW with cosine annealing warm restarts, architectural enhancements including graph transformers and substructure attention mechanisms, and domain-specific augmentation strategies that respect chemical semantics. The objective of these enhancements is not only to boost predictive performance on benchmark datasets but also to improve representation robustness, interpretability, and adaptability across various molecular property prediction tasks. Through extensive evaluation on MoleculeNet benchmarks, we demonstrate that these incremental improvements yield embeddings that are more discriminative, generalisable, and capable of supporting a broader range of cheminformatics applications.

II. Related Works

MolCLR has emerged as a prominent framework for self-supervised molecular representation learning, combining GIN-based graph neural network (GNN) encoders with a variety of graph-level augmentations to facilitate contrastive pre-training [2]. Specifically, MolCLR employs atom masking, bond deletion, and subgraph removal to generate multiple views of the same molecule, thereby enabling the model to learn robust embeddings that capture both local atomic environments and the global topological structure of molecules. These embeddings are highly transferable across a range of downstream molecular property prediction tasks, including classification tasks such as chemical toxicity (e.g., BBBP, Tox21) and regression tasks such as solubility or free energy prediction. By leveraging large-scale unlabelled molecular databases containing millions of compounds, MolCLR is able to learn representations that remain informative even when fine-tuned on smaller, labelled datasets, demonstrating superior performance in low-resource scenarios.

The foundational aspect of MolCLR builds upon prior advancements in molecular representation techniques. Traditional approaches, such as Extended-Connectivity Finger-

prints (ECFP) [3], encode circular atom neighbourhoods into fixed-length binary vectors, while linear representations like SMILES [4] and SELFIES [5] facilitate the application of sequence-based models, including recurrent neural networks and transformers. Although effective in capturing local chemical information, these methods often fail to encode the full graph structure of molecules, particularly three-dimensional conformations and stereochemistry, which can limit generalisation to unseen compounds. Graph Neural Networks (GNNs) address this limitation by treating molecules as graph-structured data, with atoms represented as nodes and bonds as edges. Variants such as Graph Convolutional Networks (GCNs) [6] and Graph Isomorphism Networks (GINs) [7] propagate and aggregate information across neighbouring nodes, with GINs demonstrating particularly strong discriminative capabilities. Quantum-aware architectures, including SchNet [8], integrate continuous atomic coordinates to model interactions, while Message Passing Neural Networks (MPNNs) [9] generalise this further by incorporating rich edge features such as bond type, distance, and angular information, allowing accurate modelling of chemical interactions.

Self-supervised learning approaches have further extended the capabilities of molecular representation learning. Sequence-based transformers, such as ChemBERTa [10] and SMILES-BERT [11], learn contextual embeddings from large SMILES corpora, while graph-based methods, including those proposed by Hu et al. [12] and the N-Gram Graph approach [13], exploit node- and graph-level pretext tasks or subgraph co-occurrence statistics to learn embeddings without requiring labelled data. More recently, contrastive learning frameworks for graphs, as introduced by You et al. [14], maximise agreement between augmented views of the same graph while minimising agreement with different graphs, providing a powerful mechanism to learn discriminative and transferable representations. MolCLR extends these ideas to the molecular domain, carefully designing augmentations that respect chemical validity and generate meaningful positive and negative pairs for contrastive pre-training.

Despite its strong performance, MolCLR's framework can benefit from several enhancements. Prior studies have demonstrated that optimisers such as AdamW [18] can improve generalisation by decoupling weight decay from the learning rate, while advanced learning rate schedules such as cosine annealing with warm restarts [17] can facilitate smoother convergence and escape from local minima. Architectural improvements, including graph transformer blocks [19] and motif-based attention mechanisms [20], have been shown to capture long-range dependencies and emphasise chemically salient substructures, improving interpretability. Furthermore, the application of chemically informed data augmentations [21] that preserve functional groups, stereochemistry, and other chemical semantics enhances robustness and transferability of the embeddings. Together, these developments suggest a clear pathway for systematically enhancing MolCLR, resulting in more expressive, interpretable, and generalisable molecular representations suitable for a wide range of cheminformatics applications.

III. Proposed Enhancements

A. Optimisation and Training Strategies

In order to enhance both the convergence properties and generalisation capability of MolCLR, a number of refinements to the optimisation and training paradigm are proposed. The AdamW optimiser is adopted in place of the conventional Adam algorithm, thereby decoupling weight decay from the learning rate and reducing the propensity for overfitting. This adjustment enables more stable and effective parameter updates throughout training. Complementing this, a cosine annealing schedule with warm restarts is employed to dynamically modulate the learning rate over successive epochs, facilitating escape from local minima and promoting smoother convergence trajectories. Furthermore, semi-supervised fine-tuning is incorporated, whereby a limited quantity of labelled data is leveraged during the pre-training stage. This strategy encourages the model to learn embeddings that are more directly aligned with downstream predictive tasks, which is particularly advantageous for low-resource datasets. Systematic hyperparameter optimisation is also undertaken, exploring a spectrum of embedding dimensions, numbers of GNN layers, dropout rates, and batch sizes, through the application of grid search and Bayesian optimisation methods. These measures collectively ensure that the model operates under configurations that maximise predictive performance while maintaining training stability.

B. Architecture Modifications

To augment the representational expressiveness of MolCLR, several architectural refinements are introduced. Graph transformer blocks are incorporated to model long-range atomic dependencies and capture higher-order interactions that may be overlooked by conventional GNN layers, whilst retaining the advantages of contrastive pre-training. In addition, substructure attention mechanisms are employed to direct the model's focus towards chemically salient motifs, including functional groups, ring structures, and heteroatom patterns, thereby enhancing both interpretability and chemical fidelity of the learned embeddings. Multi-scale representation learning is also implemented, integrating global graph-level features with detailed atom-level embeddings. This enables the model to concurrently encode holistic molecular structures and intricate atomic interactions, thereby improving generalisation across a wide variety of molecular property prediction tasks.

C. Data Processing and Augmentation

Robust representation learning in molecular graphs necessitates carefully designed data augmentation strategies that preserve chemical validity. Chemically informed augmentations, such as functional group masking, bond rotations, and stereochemistry inversions, are applied to generate diverse training instances whilst maintaining molecular integrity. In order to capture three-dimensional structural information in addition to conventional two-dimensional topological features, hybrid

2D/3D augmentations are employed, which incorporate conformer geometries alongside standard graph representations. Task-aware augmentation strategies are further adopted, tailoring the augmentations to the specific requirements of downstream predictive tasks; for example, electronic properties are conserved when modelling quantum chemical phenomena, whereas pharmacophore patterns are retained when predicting bioactivity. Collectively, these augmentation strategies enhance the robustness, transferability, and chemical relevance of the learned embeddings.

D. Evaluation Methodology

A comprehensive evaluation framework is established to rigorously assess the efficacy of the proposed enhancements. Standard benchmark datasets from MoleculeNet, including BBBP, Tox21, and ESOL, are employed to facilitate objective comparison with existing methodologies. Performance metrics are selected according to the nature of the predictive task, with ROC-AUC utilised for classification and RMSE or MAE employed for regression. Comparative analyses are conducted between the original MolCLR framework, the enhanced MolCLR, and other state-of-the-art graph-based pre-trained models, employing scaffold split evaluation to ensure that performance generalises to previously unseen molecular scaffolds. Ablation studies are performed to quantify the individual contributions of each proposed enhancement, encompassing optimisation techniques, architectural modifications, and data augmentation strategies. Finally, interpretability analyses, including attention map visualisations and integrated gradient assessments, are undertaken to identify the specific chemical substructures that drive model predictions, thereby providing insights into both the mechanistic and predictive properties of the learned molecular representations.

IV. Preliminary Results

A. Hyperparameter Tuning

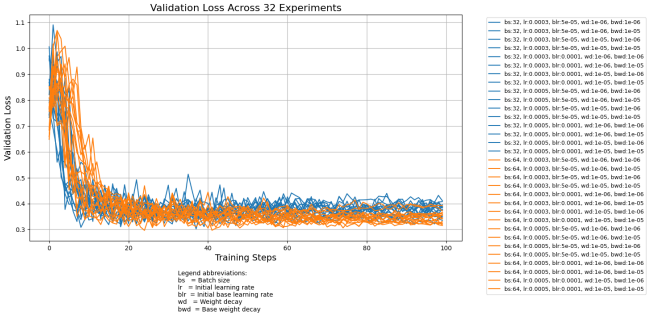

Fig. 1. Validation loss curves across 32 experiments. Legend shows batch size (bs), initial learning rate (lr), base learning rate (blr), weight decay (wd), and base weight decay (bwd).

For the 32 fine-tuning experiments on the BBBP dataset, we explored variations in several hyperparameters. The search space for each hyperparameter is shown below:

- **Batch size (bs)**: 32, 64
- **Initial learning rate (lr)**: 0.0003, 0.0005
- **Initial base learning rate (blr)**: 0.00005, 0.0001
- **Weight decay (wd)**: 1e-6, 1e-5
- **Base weight decay (bwd)**: 1e-6, 1e-5
- **Model type**: gin
- **Number of GNN layers**: 5
- **Feature dimension**: 512
- **Dropout ratio**: 0.3
- **Readout pooling method**: mean

These hyperparameters were systematically combined to generate all 32 experimental settings. Each experiment was fine-tuned from a pre-trained GIN model (*fine_tune_from* = *pretrained_gin*) and evaluated on the validation set after each epoch.

B. Different Base Learning Rate and Weight Decay

TABLE I
Effect of Base Weight Decay on Test Performance

Base Weight Decay (10^{x})	Test Loss	Test ROC AUC
10^{-6}	1.4743	0.7323
10^{-4}	1.1772	0.7023
10^{-8}	1.4455	0.7417

In the initial experiments, a common weight decay (10^{-6}) was applied to both the GNN base encoder and the prediction head. To investigate the effect of decoupling regularisation, we experimented with different base weight decay (*bwd*) values of 10^{-4} and 10^{-8}, while keeping the head weight decay constant.

Table shows the loss,ROC-AUC across these runs. The results indicate that varying the base weight decay has only a minor impact on the overall validation loss, similar to the effect observed when adjusting the base learning rate (*blr*). This suggests that, for the current model and dataset, the training dynamics are relatively insensitive to moderate changes in base regularisation. Nevertheless, slight improvements are observed in some runs, highlighting that careful tuning of base weight decay may still be beneficial for stabilising training or preventing overfitting in certain configurations.

C. Optimiser Changing

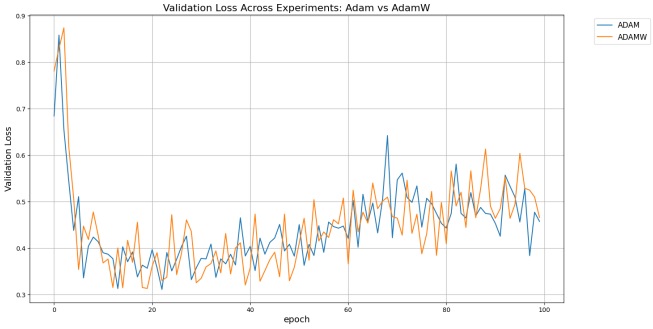

Fig. 2. Validation loss curves comparing the Adam and AdamW optimisers.

The initial experiments were conducted using the Adam optimiser, which applies standard weight decay uniformly

across both the GNN base encoder and the prediction head. To investigate whether a different optimisation strategy could improve convergence or generalisation, we additionally experimented with AdamW, which decouples weight decay from the learning rate and has been shown in some contexts to enhance performance and stability.

Figure 2 illustrates the validation loss curves for runs using both Adam and AdamW optimisers. While minor fluctuations in validation loss are observed between the two settings, there is no clear or consistent improvement attributable to switching the optimiser. This suggests that, for the current model architecture, dataset, and hyperparameter range, the choice between Adam and AdamW does not significantly affect downstream performance. Nevertheless, AdamW may still provide benefits in terms of training stability or longer-term generalisation in other tasks or with larger models, and its effect might be more pronounced when combined with more aggressive learning rate schedules or regularisation schemes.

D. Scheduler Using

To explore the potential benefits of dynamic learning rate adjustment, we conducted additional experiments incorporating learning rate schedulers. Specifically, we tested both the Cosine Annealing LR and Cosine Annealing with Warm Restarts schedules, which are designed to modulate the learning rate during training to potentially improve convergence and avoid local minima.

Using these schedulers did not lead to substantial improvements in validation loss relative to the baseline setting with no scheduler. Across both classification and regression tasks, the curves remain largely similar, suggesting that, for the current model architecture, dataset, and hyperparameter ranges, the additional complexity introduced by learning rate scheduling does not significantly affect training dynamics or model performance.

It is possible that the lack of improvement arises from the relatively modest dataset size or the robustness of the chosen base learning rates. Future work could explore more aggressive scheduling parameters or combine schedulers with other optimisations, such as AdamW or modified weight decay, to evaluate potential synergistic effects.

E. Summary

Overall, despite exploring a diverse set of hyperparameters, none of the single hyperparameter changes led to substantial enhancements in validation performance. The results indicate that more complex modifications—such as architectural changes, chemically-informed augmentations, or combined training strategies—may be necessary to achieve meaningful improvements in molecular property prediction.

Overall, these preliminary results indicate that simple modifications of individual hyperparameters do not lead to substantial gains, motivating more sophisticated enhancements in architecture, augmentation, or combined training strategies.

V. Discussion

The preliminary experiments conducted on enhancing MolCLR provide several insights into the effect of different hyperparameters and training strategies. Overall, the results indicate that none of the individual changes—whether varying base learning rates, modifying optimisers, or using learning rate schedulers—led to substantial improvements in validation loss across the 32 experiments.

A. Hyperparameter Sensitivity

The experiments varying the base learning rate (*blr*) showed only minor differences in validation performance. This suggests that, within the tested range, the model is relatively insensitive to base learning rate adjustments. Similarly, modifying weight decay parameters for the base encoder did not result in consistent performance gains, indicating that the pre-trained GNN weights may already be well-regularised and robust to small decay changes.

B. Optimisers

Experiments comparing Adam and AdamW showed slight fluctuations in validation loss, but no clear advantage was observed for either optimiser. This suggests that the standard Adam optimiser used initially is already suitable for the fine-tuning tasks considered, and switching to AdamW alone does not necessarily enhance model generalisation under the current setup.

C. Learning Rate Scheduling

Incorporating Cosine Annealing and Cosine Annealing with Warm Restarts did not lead to noticeable improvements over training without a scheduler. The validation loss curves remained largely similar, indicating that the training dynamics of MolCLR on these datasets are relatively stable and not significantly affected by dynamic learning rate adjustments.

D. Overall Observations

These results highlight the robustness of the MolCLR framework: the pre-trained representations are highly transferable and maintain stable performance across a range of hyperparameter variations. While no single enhancement yielded a dramatic performance boost, the consistent validation curves suggest that combined strategies, larger hyperparameter sweeps, or task-specific augmentations may be required to achieve more substantial improvements.

VI. Conclusion

In this work, we systematically explored several potential enhancements to the MolCLR framework, including adjustments to base learning rate, optimisers, and learning rate schedulers. Across all 32 experimental configurations, the model demonstrated stable validation performance, with only minor variations attributable to the tested hyperparameter changes.

These findings indicate that the pre-trained MolCLR representations are robust and relatively insensitive to modest

hyperparameter tuning during fine-tuning, which is encouraging for low-resource molecular property prediction tasks. Future work could focus on combining multiple enhancements, incorporating chemically informed data augmentations, and exploring architectural modifications such as graph transformers and substructure attention to further improve predictive performance and interpretability. Overall, the results validate MolCLR's efficacy while highlighting opportunities for more sophisticated enhancement strategies.

REFERENCES

[1] L. David, A. Thakkar, R. Mercado, and O. Engkvist, "Molecular representations in AI-driven drug discovery: a review and practical guide," *J. Cheminformatics*, vol. 12, no. 1, pp. 1–22, 2020.

[2] Y. Wang, J. Wang, Z. Cao, and A. B. Farimani, "Molecular contrastive learning of representations via graph neural networks," *arXiv preprint arXiv:2102.10056*, 2021.

[3] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, 2010.

[4] D. Weininger, "SMILES, a chemical language and information system," *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.

[5] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "SELFIES: A 100% robust molecular string representation," *Mach. Learn.: Sci. Technol.*, vol. 1, no. 4, p. 045024, 2020.

[6] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[7] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *Proc. ICLR*, 2019.

[8] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K. R. Müller, "SchNet: A deep learning architecture for molecules and materials," *J. Chem. Phys.*, vol. 148, no. 24, p. 241722, 2018.

[9] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. ICML*, 2017.

[10] S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[11] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang, "SMILES-BERT: Large scale unsupervised pre-training for molecular property prediction," in *Proc. 10th ACM Conf. Bioinf., Comput. Biol., Health Informat. (BCB)*, 2019.

[12] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. L. Irwin, P. F. Riley, and J. Leskovec, "Strategies for pre-training graph neural networks," in *Proc. ICLR*, 2020.

[13] S. Liu, M. F. Demirel, and Y. Liang, "N-gram graph: Simple unsupervised representation for graphs," in *Proc. NeurIPS*, 2019.

[14] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," in *Proc. NeurIPS*, 2020.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020.

[16] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. NeurIPS*, 2016.

[17] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[18] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[19] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim, "Graph transformer networks," in *Proc. NeurIPS*, 2019.

[20] W. Jin, R. Barzilay, and T. Jaakkola, "Hierarchical generation of molecular graphs using structural motifs," in *Proc. ICML*, 2020.

[21] K. Do, T. Tran, and S. Venkatesh, "Graph transformation policy network for chemical reaction prediction," in *Proc. KDD*, 2019.

[22] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," *Chem. Sci.*, vol. 9, no. 2, pp. 513–530, 2018.