Developing novel augmentation strategies for contrastive learning in molecular property prediction

Duc Ho, Viet Chu, Farel Lukas, Atharva Tamore,
Aditya Vaidya, Khanh Bui, Darlene Nguyen, Xinje Ye
Department of Computing Science, University of Alberta
{dtho, vchu1, farelnic, atamore, avaidya1, kdbui, maianh1, xye1}@ualberta.ca

Abstract

Graph neural networks (GNN) combined with contrastive learning have emerged as a promising machine learning tool for predicting molecular properties. However, their effectiveness heavily depends on the choice of molecular graph augmentation techniques. In this project, we explore 8 different innovative augmentation strategies within the MolCLR framework and evaluate them on 3 molecular property prediction tasks: BACE, BBBP, and HIV. Our experiments focused on the curated Myopic MCES dataset with 20,000 molecules to assess the effectiveness of each augmentation method individually and in combination. While our results did not show significant improvement in performance compared with the original techniques that MolCLR used, the consistency of our chemical-based augmentations suggested that chemical knowledge awareness can enhance the performance of the MolCLR pipeline. Our results also suggested that MolCLR can utilize a significantly smaller dataset to achieve similar performance, saving computational and time resources.

1 Introduction

Molecular property prediction plays a vital role in drug discovery and computational chemistry by enabling the assessment of compound properties without costly lab experiments. Graph Neural Networks (GNNs) have advanced this field by modeling molecules as graphs, with atoms as nodes and bonds as edges [4, 5]. However, supervised GNN models rely on large labeled datasets, which are scarce due to the expense of wet-lab data collection.

Contrastive learning offers a self-supervised alternative that learns meaningful molecular representations by contrasting different augmented views of the same molecule [10]. The success of this method heavily depends on the augmentation strategies used, yet how these techniques apply to molecular graphs—especially with chemical context in mind—remains underexplored [3, 11].

This project investigates chemically-informed augmentation strategies within the MolCLR framework [10]. We experiment with methods like functional group-aware masking and alkylation-based perturbations, evaluating their effects on molecular representation learning across three MoleculeNet tasks—BACE, BBBP, and HIV—using datasets from the Myopic MCES repository [7]. Our findings show that certain augmentations consistently lead to better model performance, underscoring the value of domain knowledge in molecular contrastive learning.

2 Background

2.1 MolCLR Framework

MolCLR applies a self-supervised contrastive learning strategy specifically adapted for graph data to predict the molecular properties. This framework utilizes graph neural networks (GNN) to generate embeddings from molecular graphs. To create diverse representations, MolCLR uses three main augmentation techniques - atom masking, bond deletion, and subgraph removal. Their research has shown the potential of MolCLR in enhancing the accuracy of molecular property predictions when compared to standard supervised learning methods Wang u.a. [10] (see 10 in the Appendix).

2.2 Contrastive Learning

Contrastive learning is a self-supervised strategy for producing robust data representations. The idea is to teach a model to identify similarities and differences by contrasting related data points (positive pairs) with unrelated ones (negative pairs). By learning to group variations of the same data together and separate different data points, the model develops generalizable embeddings that can be effectively used for tasks, such as classification and regression [11].

2.3 Graph Data Augmentation

Graph data augmentation refers to strategies that create variant graph representations by applying modifications to node features, edge structures, or graph topology. According to Ding u.a. [3], augmentations for deep graph learning can be categorized into feature-oriented (e.g., node masking), structure-oriented (e.g., edge deletion), and labeled-oriented (e.g., label mixing) augmentations, serving different purposes. The survey emphasizes that the design of augmentations should align with the nature of the graph task. For molecular graphs of this project, in particular, this means preserving chemical validity while introducing meaningful variance. Augmentation quality significantly impacts the model's performance.

3 Related Works

Many prior studies have successfully applied contrastive learning techniques to molecular graph prediction tasks, demonstrating certain performance improvements through innovative augmentation methods such as those used in GraphCL [11, 8], and structural integrity-preserving perturbations by [6]. Our augmentation approaches drew inspiration from:

1. **MolNexTR**: Focused on addressing noise and visual clutter in chemical structure recognition via "Image Contamination". We adopted this idea in our Alkylation augmentation by adding structured noise to molecules Chen u.a. [2].
2. **Data-centric Properties**: Emphasized invariance, separability, and recoverability as principles for designing meaningful augmentations [9]. We used these principles to build our Functional Group augmentations, where we utilized essential chemical information in our augmentations.
3. **MoCL**: Showcased that augmentations informed by chemical knowledge outperform random ones, motivating our focus on Functional Group augmentations [8].

4 Methodology

4.1 Data

We used the Myopic MCES dataset, consisting of 20K unique molecular graphs from Kretschmer u.a. [7]. For downstream tasks, we chose three standard molecular classification benchmarks from MoleculeNet: BACE, BBBP, and HIV. These are recognized as crucial datasets for determining the effectiveness of molecular machine learning models [10]. During pretraining, we use 95% of the dataset for training, and 5% for validation after each epoch of training the MolCLR. For finetuning, 80% of the dataset goes into training, 10% for validation after each epoch, and 10% is reserved for testing.

4.2 Pipeline

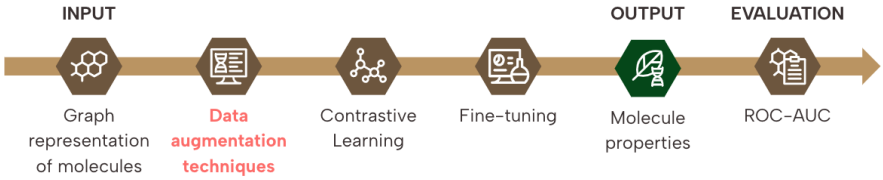


Figure 1: Overview of our pipeline for molecular property predictions using contrastive learning.

We adapted the MolCLR pipeline for contrastive pretraining and downstream tasks, with modifications only in the augmentation stage. The pipeline consists of the following steps:

1. **Input**: Molecular structures from the Myopic MCES dataset are provided in SMILES format and constructed into graph representations.
2. **Data Augmentation Techniques**: Each input molecular graph is transformed into two distinct augmented representations using one or more augmentation techniques.
3. **Contrastive Learning (pretraining)**: Each augmented graph is encoded by a shared GNN and mapped into a contrastive embedding space for training with NT-Xent contrastive loss, to align positive pairs and separate negatives.
4. **Finetuning**: After pretraining, the GNN encoder is finetuned on labeled datasets for a specific downstream task. During this step, the model finetunes its parameters based on labeled molecular graphs to improve its generalization.
5. **Output and Evaluation**: The output of the finetuned model consists of predicted labels associated with a specific molecular property. Evaluation is performed using ROC-AUC scores on benchmark classification datasets to assess model performance.

4.3 Augmentations

To see the impact of data augmentation techniques on the contrastive learning setup for molecular property prediction, we tried out

4.3.1 MolCLR original augmentations

Wang u.a. [10] originally used 3 augmentations in their MolCLR pipeline:

1. **Atom Masking and Bond Deletion**:
   - Atom Masking randomly masks some atoms in the molecule graph with a given ratio. Every masked atom has its feature x_v replaced by a mask token, m, which is different from any atom features in the molecule graph. This is shown by the red box in figure 2. Masking forces the model to learn the intrinsic chemical information within molecules, such as possible types of atoms connected by certain covalent bonds [10].
   - Bond Deletion randomly removes chemical bonds between atoms at a given ratio, shown in the yellow box in Figure 2. Unlike Atom Masking, it alters the graph structure by breaking edges, simulating real chemical bond breakage and encouraging the model to learn how molecules behave in reactions.
   - Wang u.a. [10] combined these augmentations in their experiment by first applying Atom Masking, then Bond Deletion, on the same molecule, with the ratio set to 25% for both augmentations. This method resulted in 2 molecular graphs, each graph being both atom-masked and bond-deleted. In the later section, we will introduce a different integration of these two: rather than having a molecule graph go through both of them, we duplicate the graph into two graphs; one will go through Atom Masking, and the other will go through Bond Deletion.

2. **Subgraph Removal**: combines Atom Masking and Bond Deletion by masking atoms and deleting their bonds in a breadth-first manner from a randomly selected origin, stopping once a target ratio of masked atoms is reached. As shown in the blue box in Figure 2, this helps the model learn key structural motifs by comparing subgraph-removed variants [10]. Mixing applies both augmentations sequentially. A subgraph is first removed with a random ratio (0–25%). If the masked atom or bond deletion ratio is below 25%, additional atoms or bonds are randomly removed to meet the target.

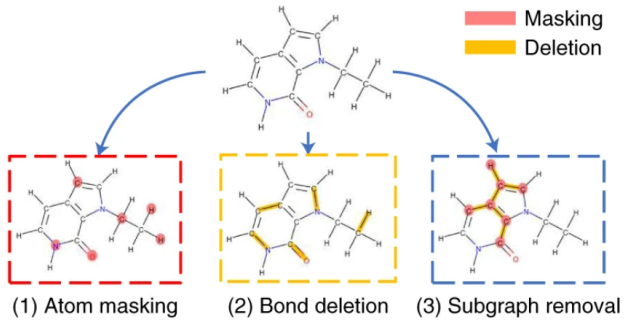
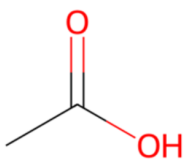


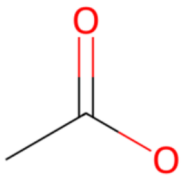
Figure 2: Three augmentations used by Wang u.a. [10] in their MolCLR pipeline.

4.3.2 Atom Masking and Bond Deletion 2

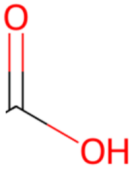
Using the inspiration from You u.a. [11], we re-designed the combination of Atom Masking and Bond Deletion using a simple idea: instead of applying node and edge augmentations simultaneously; we generate two distinct views of the same molecule - one with dropped nodes and the other with dropped edges. Unlike the original combination that apply multiple augmentations to a single molecular graph, often leading to excessive corruption and loss of chemical information, our method applies only one type of structural change to each graph. The goal was to make each view look structurally different while still representing the same molecule, allowing the model to learn more meaningful similarities. This new strategy encouraged the model to focus more on the core structure of the molecule.



(a) An example molecule graph



(b) First duplicated graph with Atom Masking. A Hydrogen atom has been removed



(c) Second duplicated graph with Bond Deletion. A Carbon bond has been removed

Figure 3: An example of Atom Masking and Bond Deletion 2. Two augmented graphs are produced

4.3.3 Edge Perturbation

We modified the MolCLR framework to introduce an edge augmentation method called Edge Perturbation, which adds a third, continuous feature channel to standard 2D edge attributes (bond type and direction). This new channel consists of random values from a Gaussian distribution, scaled by a hyperparameter `perturbation_scale`. GIN and GCN models were adapted with MLP-based edge encoders to process these 3D edge features, learning embeddings that blend discrete bond information with random noise. This trains the model to focus on stable structural features rather than the noise. Empirically, this simple augmentation outperformed more complex spectral-based perturbations [6] during finetuning.

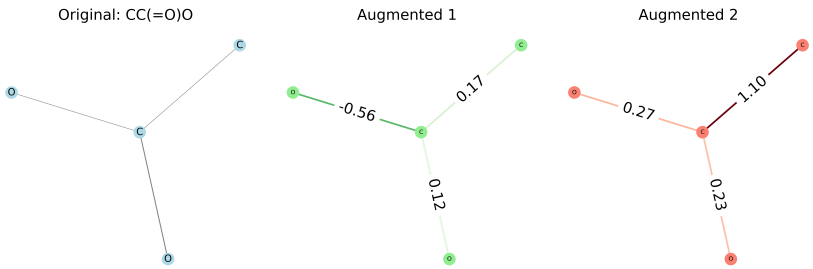


Figure 4: Edge perturbation applied on two augmented pairs of an example molecule. The numbers on the edge are called edge weights, which are sampled from $\mathcal{N}(0, 1)$.

4.3.4 Functional Group Augmentation

Functional groups (FG) are specific combinations of atoms within a molecule that determine its chemical and physical properties, as well as its characteristic reactivity patterns. Because they carry important chemical meaning, we focused on them in our following augmentations to help the model better capture key structural and functional patterns.

1. **Functional Group Exclusion (FG Exclusion)**: Inspired by MolNexTR's abbreviation rules, we selectively exclude atoms from MolCLR's atom masking augmentation if they belong to recognized functional groups. This structural-preserving approach prevents masking of these critical units, ensuring that the model retains essential chemical information during training.

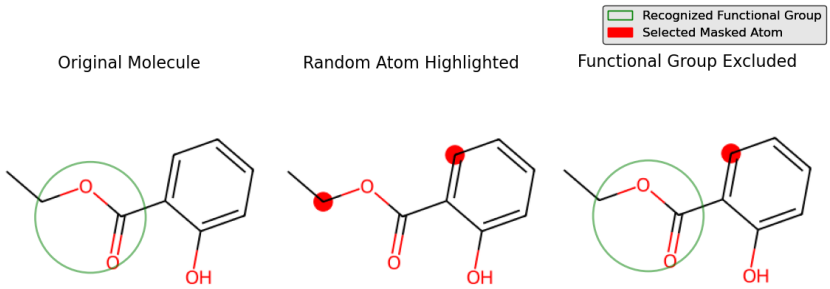


Figure 5: Functional Group Exclusion example

2. **Functional Group Mixed-masking (FG Mixing)**: This technique applies three different masking approaches with equal probability. Specifically, 1/3 of the time, the augmentation randomly masks nodes across the graph; 1/3 of the time, it masks the entire functional group; and the remaining 1/3 of the time, it excludes the functional group entirely from being masked. This approach is designed to balance the introduction of variability with the preservation of essential chemical information.

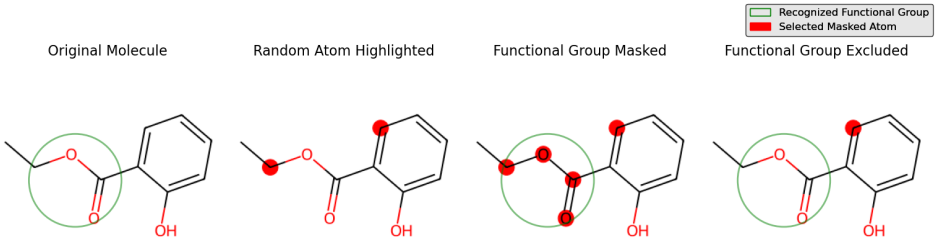


Figure 6: Functional Group Mixed-masking example.

3. **Functional Group Replacement (FG Replacement)**: functional groups within molecules were replaced with chemically equivalent or similar substructures, guided by bioisosteric

substitution rules. This augmentation maintained chemical validity and relevance, following the methodologies proposed in Sun et al. (2021).

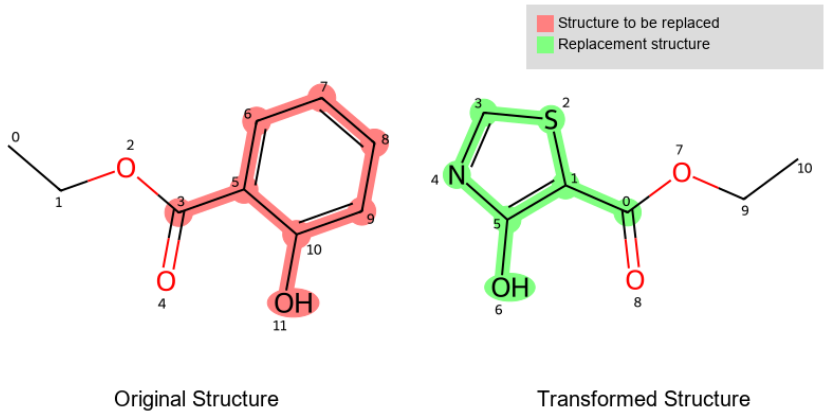


Figure 7: Functional Group Replacement example

4.3.5 Alkylation

We introduced an alkylation-based node addition augmentation as another novel augmentation strategy. Essentially, we mimic an organic alkylation reaction by randomly selecting an atom in the molecular graph and appending a new carbon node through a single edge. Our hope is that the model can generalize across molecules sharing the same functional groups since these groups often exhibit related behaviors in chemistry. In most cases, these controlled modifications only adjust chemical properties slightly, encouraging the model to focus on core structural patterns. Of course, in the real world, alkylation can sometimes cause more noticeable shifts, such as notable pH changes in certain homologous series, so we are mindful not to overgeneralize [1].

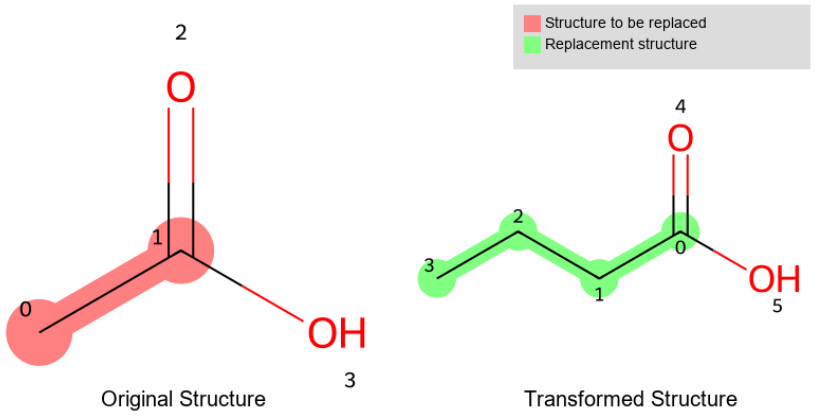


Figure 8: Three augmentations used by Wang u.a. [10] in their MolCLR pipeline.

5 Results

5.1 Experiment Setting

Using the 20K dataset for pretraining, we finetuned the models five times on each downstream task using the configuration from the table 3 in the Appendix, but with different pretraining augmentations.

5.2 Experiment Results

Augmentation	BACE	BBBP	HIV
No Augmentation	77.37 (± 2.08)	66.20 (± 2.23)	74.01 (± 1.58)
Atom Masking & Bond Deletion 1	**82.88 (± 1.79)**	70.38 (± 1.64)	75.42 (± 1.54)
Subgraph	82.35 (± 0.46)	**72.59 (± 0.50)**	76.02 (± 1.25)
Mixing	80.62 (± 0.92)	70.68 (± 2.82)	75.62 (± 1.61)
Atom Masking & Bond Deletion 2	79.72 (± 2.22)	68.55 (± 2.56)	**76.40 (± 0.65)**
Alkylation	79.32 (± 0.98)	67.73 (± 2.04)	76.19 (± 0.93)
Edge Perturbation GIN	80.66 (± 2.59)	66.92 (± 0.90)	75.93 (± 1.62)
Edge Perturbation GCN	74.78 (± 1.12)	64.49 (± 1.55)	73.50 (± 0.65)
FG Mixing & Bond Deletion	81.91 (± 0.91)	71.59 (± 1.39)	76.23 (± 1.00)
FG Exclusion & Bond Deletion	80.27 (± 1.50)	69.36 (± 0.86)	76.06 (± 1.12)
FG Replacement	81.01 (± 1.53)	71.90 (± 1.25)	75.84 (± 1.89)

Table 1: Test performance of different augmentations on 3 downstream tasks with 20K pretraining dataset. Mean and standard deviation of test ROC-AUC (%) on each benchmark are reported. Bold augmentations are from the MolCLR paper.

Even though our models are trained on a substantially smaller dataset (20K Myopic MCES vs. 10M PubChem), we are able to achieve performance that is remarkably close to the original MolCLR results. Using the same model architecture and augmentation strategies, our recreated framework yields competitive ROC-AUC scores across BBBP, BACE, and HIV benchmarks. This highlights the strong representational power of our curated dataset and the potential for effective pretraining even with limited data volume.

Further evaluations showed that with regard to no augmentations, nearly all augmentations improved the performance of the model on the 3 downstream tasks, except for Edge Perturbation GCN. Edge Perturbation GCN performed poorly because it uses GCN for its backbone, while all other augmentations, including "No Augmentation", used GIN, which has been shown to be better in most cases [10]. All augmentations performed somewhat similarly to each other in each downstream tasks, except for BBBP, where Atom Masking & Bond Deletion 2, Alkylation and Edge Perturbation lost entirely compared to the MolCLR original augmentations. This could be the result of the topological sensitivity of BBBP molecules, where a small perturbation in the structures of the molecular graph could result in a large change in chemical properties [10].

We also ran some experiments with the Function Group Replacement augmentation on the 740K dataset and produced the following results.

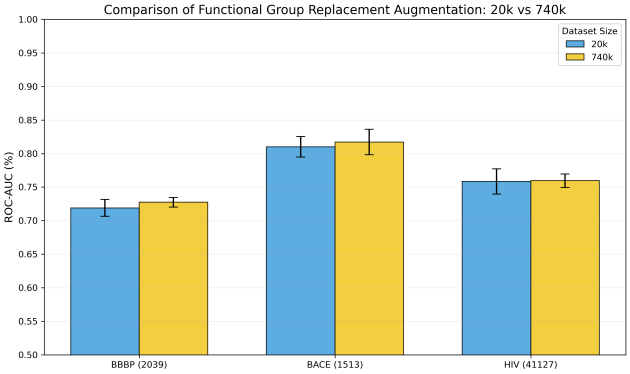


Figure 9: Test performance of Function Group Replacement on 3 downstream tasks with both 20K pretraining dataset and 740K pretraining dataset. The height of each bar represents the mean ROC-AUC (%) on the downstream task, and the length of each error bar represents the standard deviation.

These results showed that even though we provided more data for pretraining, finetuning performance did not improve significantly.

We conducted focused comparisons between functional group-based augmentations and their structurally related counterparts — specifically, atom-level masking vs. functional group mixing, and subgraph masking vs. functional group replacement. These pairings were chosen because functional group mixing is a knowledge-informed variation of random atom masking, while functional group replacement mirrors the structure-level modifications in subgraph masking. Our goal was to assess how domain knowledge influences representation learning.

Results show that functional group mixing slightly outperforms atom-level masking on BBBP and HIV, indicating that even modest knowledge-based changes can help (see Figure 15 in the Appendix). In contrast, functional group replacement under-performs subgraph masking, despite prior success reported by Sun u.a. [8]. We suspect this is due to differing pretraining setups: their models were pretrained directly on downstream molecules, making domain knowledge more aligned with task-specific features, while our pretraining uses a general dataset.

6 Discussion

Augmentations significantly outperformed non-augmentation in contrastive learning, but no single method stood out across tasks, suggesting that model architecture, not augmentation, may be the key to further improvement. Despite using much smaller datasets (20K and 740K) than the 10M used in Wang u.a. [10], performance remained comparable, indicating limited benefit from the larger dataset. This also highlights the effectiveness of the sub-sampling method from Kretschmer u.a. [7], which ensured strong data coverage even with a small subset.

One crucial thing that this project did not address is statistical testings for the difference in performance of the augmentations. Due to computational and time limitations, the number of finetuning experiment we conducted is restricted, leading to not having enough data that we could use to find any significant improvement of these augmentations.

Another thing our results suggested is chemical-based augmentations (the Functional Group augmentations) might performed better with sensitive data, such as the BBBP dataset, than image-based (Alkylation) and pure graph augmentations (Edge Perturbation).

7 Conclusion

In conclusion, we explored the impact of data augmentation on contrastive learning for molecular property prediction using a 20K high-coverage dataset from Kretschmer u.a. [7]. Testing 11 augmentations (including 3 from MolCLR) across BACE, BBBP, and HIV tasks, we found similar performance across most methods, except Alkylation and Edge Perturbation, which underperformed on BBBP. Pretraining on a larger 740K dataset showed no significant improvement over the 20K set, suggesting either insufficient scale or strong coverage by the smaller dataset.

7.1 Future works

In the future, we might conduct more experiments for each of the augmentations in each downstream task to gather enough data for statistical tests. We also want to apply more chemical-based augmentations to confirm our hypothesis about the better performance yielded by the chemical-based techniques over the non-chemical ones.

8 Acknowledgment

We would like to sincerely thank Dr. Russ Greiner for his invaluable guidance throughout this project. We also extend our gratitude to our Teaching Assistant, Weijie, for his support, and to our domain experts, Sajad Ramezan and Fei Wang, for their insightful feedback and expertise.

References

[1] Ball, David W. / Hill, John W. / Scott, Rhonda J. (2011): *The Basics of General, Organic, and Biological Chemistry*. , Flat World Knowledge.

[2] Chen, Yufan / Leung, Ching Ting / Huang, Yong / Sun, Jianwei / Chen, Hao / Gao, Hanyu (2024): *MolNexTR: a generalized deep learning model for molecular image recognition* .

[3] Ding, Kaize / Xu, Zhe / Tong, Hanghang / Liu, Huan (2022): *Data Augmentation for Deep Graph Learning: A Survey* .

[4] Duvenaud, David / Maclaurin, Dougal / Aguilera Iparraguirre, Jorge / Gómez Bombarelli, Rafael / Hirzel, Timothy / Aspuru Guzik, Alán / Adams, Ryan P. (2015): *Convolutional Networks on Graphs for Learning Molecular Fingerprints* .

[5] Gilmer, Justin / Schoenholz, Samuel S. / Riley, Patrick F. / Vinyals, Oriol / Dahl, George E. (2017): *Neural Message Passing for Quantum Chemistry* .

[6] Ko, Taewook / Choi, Yoonhyuk / Kim, Chong Kwon (2023): *Universal Graph Contrastive Learning with a Novel Laplacian Perturbation* .

[7] Kretschmer, Fleming / Seipp, Jan / Ludwig, Marcus / Klau, Gunnar W. / Böcker, Sebastian (2025): *Coverage bias in small molecule machine learning* .

[8] Sun, Mengying / Xing, Jing / Wang, Huijun / Chen, Bin / Zhou, Jiayu (2022): *MoCL: Data-driven Molecular Fingerprint via Knowledge-aware Contrastive Learning from Molecular Graph* .

[9] Trivedi, Puja / Lubana, Ekdeep Singh / Heimann, Mark / Koutra, Danai / Thiagarajan, Jayaraman J. (2023): *Analyzing Data-Centric Properties for Graph Contrastive Learning* .

[10] Wang, Yuyang / Wang, Jianren / Cao, Zhonglin / Barati Farimani, Amir (2022): *Molecular contrastive learning of representations via graph neural networks* .

[11] You, Yuning / Chen, Tianlong / Sui, Yongduo / Chen, Ting / Wang, Zhangyang / Shen, Yang (2021): *Graph Contrastive Learning with Augmentations* .

A Appendix / supplemental material

Dataset	# Molecules	# Tasks	Task type	Metric	Split
BBBP	2039	1	Classification	ROC-AUC	Scaffold
HIV	41127	1	Classification	ROC-AUC	Scaffold
BACE	1513	1	Classification	ROC-AUC	Scaffold

Table 2: Summary of three downstream task dataset for molecular property predictions.

Hyperparameter	**MolCLR Paper**	**T5-OptML**
Pretraining Dataset	PubChem (∼10M molecules)	myopic MCES (20K molecules subset)
Finetuning Datasets	BACE, BBBP, ClinTox, MUV, etc.	BACE, BBBP, HIV
Metric	ROC-AUC	ROC-AUC
Pooling Method	Average Pooling	Average Pooling
Loss Function	NT-Xent Contrastive Loss	NT-Xent Contrastive Loss
Batch Size	512	512
Number of GNN Layers	5	5
Epochs (pretraining)	100	100
Epochs (finetuning)	100	100
Embedding Dimension	300	300
Feature Dimension	512	512

Table 3: Comparison of hyperparameters between MolCLR and this project

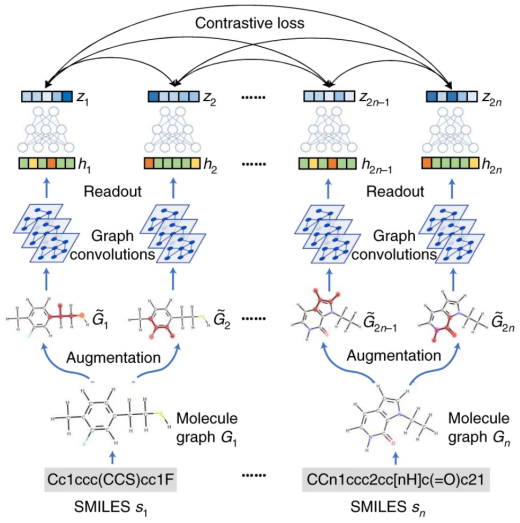


Figure 10: MolCLR architecture from Wang u.a. [10]. A SMILES string s_n is turned into graph representation and augmented into 2 correlated masked graph. These graphs are then feed into a series of graph convolution layers that produced two representation vector $h_2n - 1$ and h_2n. Contrastive loss are utilized to minimize the disagreement between the latent vector $z_2n - 1$ and z_2n from the MLP projection head.

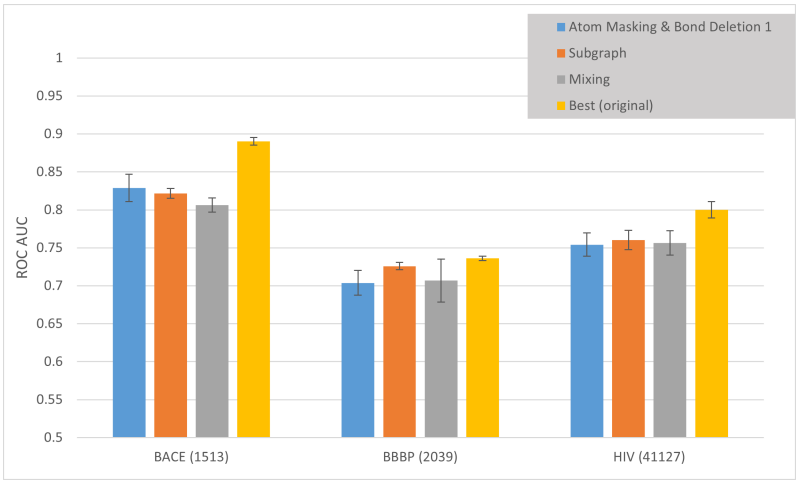


Figure 11: Comparison of ROC-AUC performance between models trained on the 20K Myopic MCES dataset and those trained on the 10M PubChem dataset using MolCLR augmentation strategies.

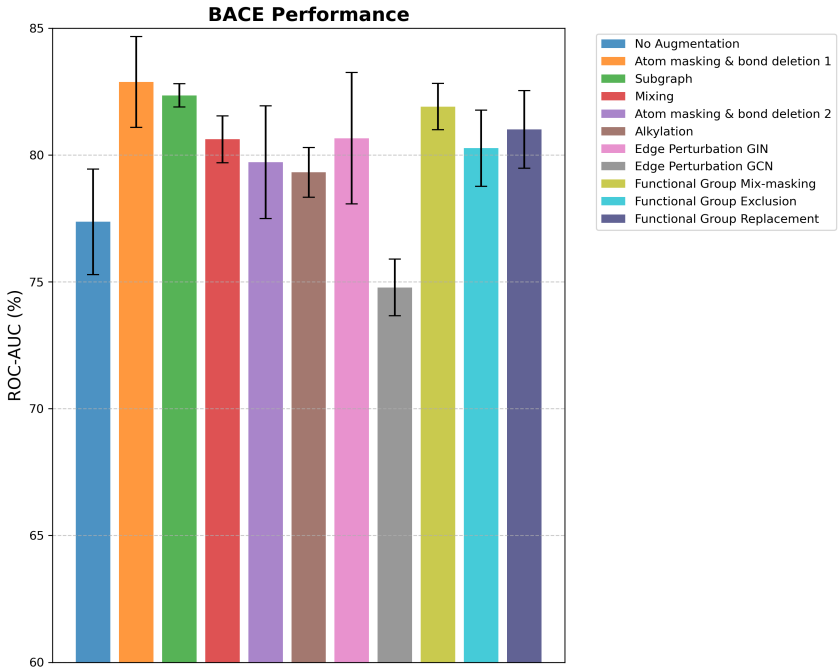


Figure 12: BACE performance on all augmentation on 20k dataset

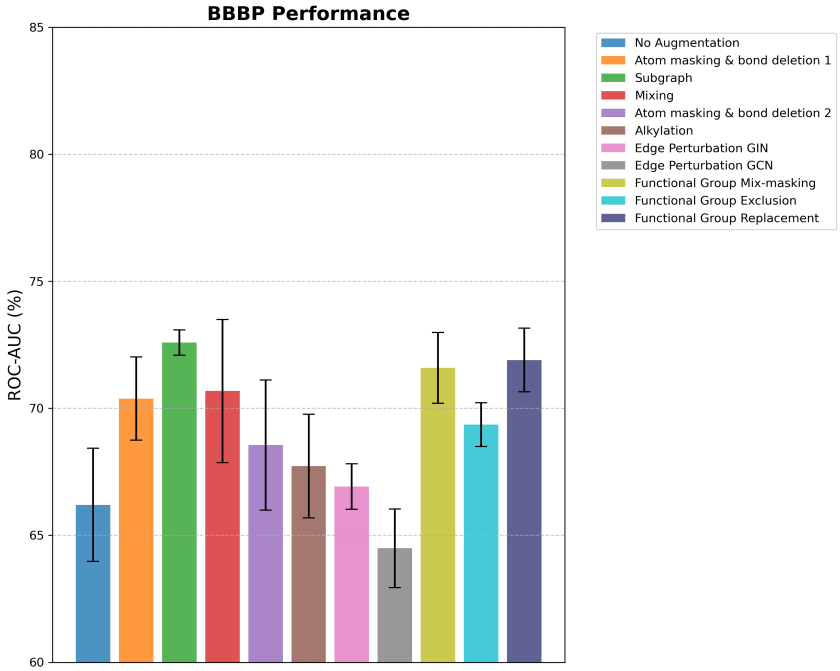


Figure 13: BBBP performance on all augmentation on 20k dataset

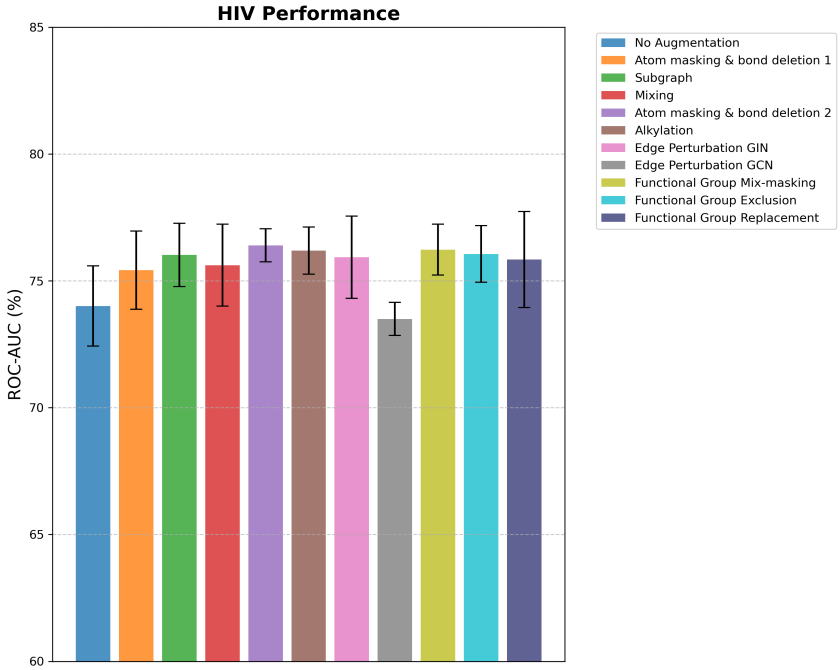


Figure 14: HIV performance on all augmentation on 20k dataset

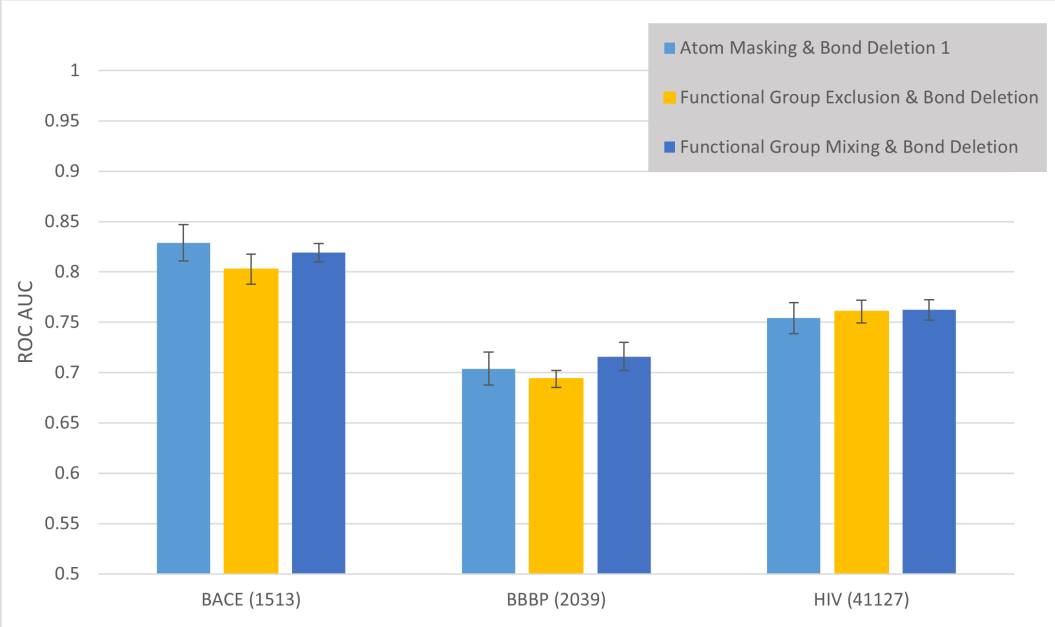


Figure 15: Atom-level vs. Functional group-level representations.

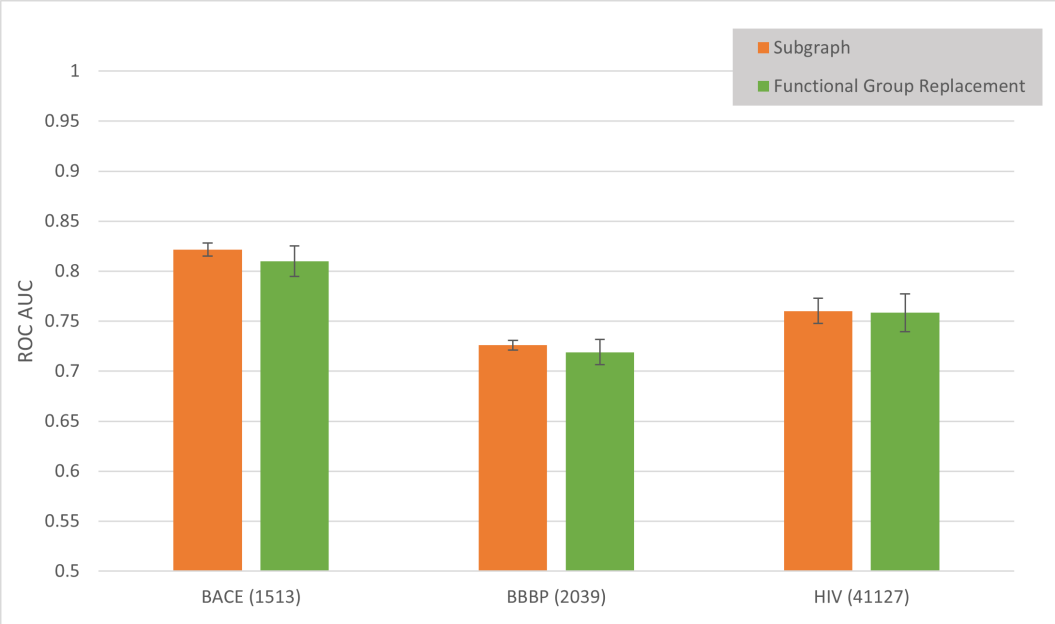


Figure 16: Subgraph removal vs. Functional group replacement.