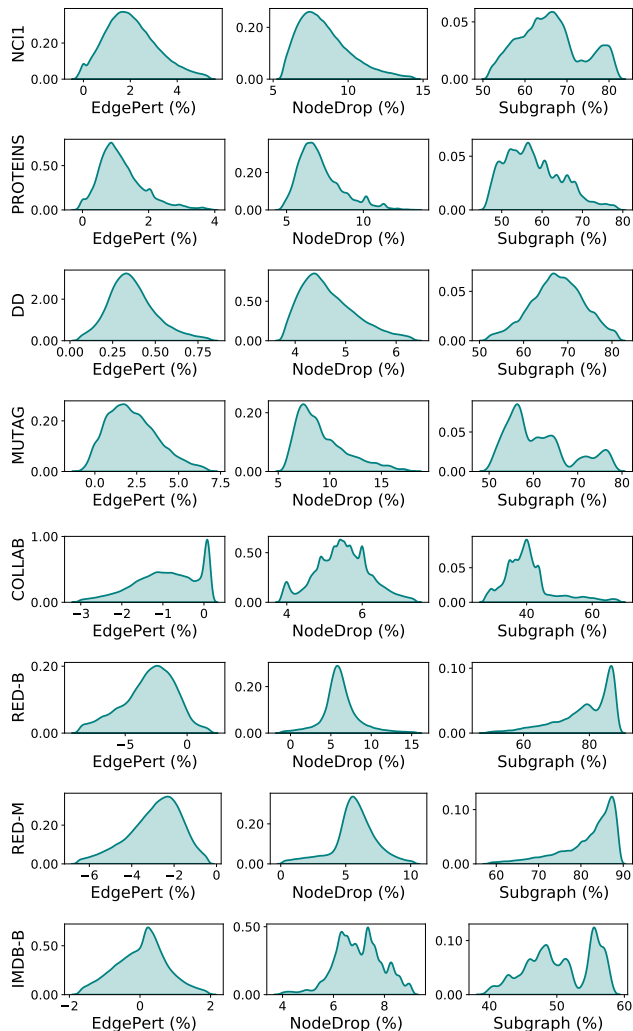


# Supplementary Materials: Uncovering Capabilities of Model Pruning in Graph Contrastive Learning

Anonymous Authors

## A QUANTIFICATION OF STRUCTURAL DAMAGE FROM DATA AUGMENTATION



**Figure 1: Quantification of structural damage from data augmentation. Percent change in structural entropy after data augmentation (i.e., Edge perturbation, Node dropping, and Subgraph with 20% strength from GraphCL).**

**Data Augmentations on Graphs.** Follow the data augmentations in GraphCL [22], we adopt three types of general data augmentations for graph-structured data:

- **Node dropping.** Given the graph  $G$ , node dropping will randomly discard certain portion of vertices along with their

connections. The underlying prior enforced by it is that missing part of vertices does not affect the semantic meaning of  $G$ . Each node’s dropping probability follows a default i.i.d. uniform distribution (or any other distribution).

- **Edge perturbation.** It will perturb the connectivities in  $G$  through randomly adding or dropping certain ratio of edges. It implies that the semantic meaning of  $G$  has certain robustness to the edge connectivity pattern variances. We also follow an i.i.d. uniform distribution to drop each edge.
- **Subgraph.** This one samples a subgraph from  $G$  using random walk. It assumes that the semantics of  $G$  can be much preserved in its (partial) local structure.

The quantitative illustrations of structural damage caused by three data augmentation rules on eight datasets are shown in Figure 1. As can be seen, the effect of structural damage varies with the augmentation rules. Specifically, node dropping and subgraph lead to different degrees of structural damage, and the information loss caused by subgraph is the largest and generally over 50%. Besides the simple information loss, the structure damage composition of edge perturbation is more complex; put differently, edge perturbation even introduces external data noise with the additional edges, which further interferes with the model from learning the actual structural information.

## B SUMMARY OF DATASETS

### B.1 Datasets for Unsupervised Learning

A wide variety of datasets from different domains for a range of graph property prediction tasks are used for our experiments. Here, we present detailed descriptions of the 8 benchmarks utilized in this paper. Table 1 shows statistics for datasets.

**Social Network Datasets.** IMDB-BINARY is derived from the collaboration of a movie set. In this dataset, every graph consists of actors or actresses, and each edge between two nodes represents their cooperation in a certain movie. Each graph is derived from a prespecified movie, and its label corresponds to the genre of this movie. Similarly, COLLAB is also a collaboration dataset but from a scientific realm, which includes three public collaboration datasets (i.e., Astro Physics, High Energy Physics and Condensed Matter Physics). Many researchers from each field form various ego networks for the graphs in this benchmark. The label of each graph is the research field to which the nodes belong. REDDIT-BINARY and REDDIT-MULTI-5K are balanced datasets, where each graph corresponds to an online discussion thread and nodes correspond to users. An edge is drawn between two nodes if at least one of them responds to another’s comment. The task is to classify each graph into the community or subreddit to which it belongs.

**Small Molecules.** NCI1 is a dataset made publicly available by the National Cancer Institute (NCI) and is a subset of balanced datasets containing chemical compounds screened for their ability

**Table 1: Statistics for datasets of diverse nature from the benchmark TUDataset.**

| Dataset         | #Graphs | #Classes | Avg. #Nodes | Avg. #Edges |
|-----------------|---------|----------|-------------|-------------|
| Social Networks |         |          |             |             |
| COLLAB          | 5,000   | 3        | 74.49       | 2457.78     |
| REDDIT-BINARY   | 2,000   | 2        | 429.63      | 497.75      |
| REDDIT-MULTI-5K | 4,999   | 5        | 508.52      | 594.87      |
| IMDB-BINARY     | 1,000   | 2        | 19.77       | 96.53       |
| Small Molecules |         |          |             |             |
| NCI1            | 4,110   | 2        | 29.87       | 32.30       |
| MUTAG           | 188     | 2        | 17.93       | 19.79       |
| Bioinformatics  |         |          |             |             |
| PROTEINS        | 1,113   | 2        | 39.06       | 72.82       |
| DD              | 1,178   | 2        | 284.32      | 715.66      |

to suppress or inhibit the growth of a panel of human tumor cell lines; this dataset possesses 37 discrete labels. MUTAG has seven kinds of graphs that are derived from 188 mutagenic aromatic and heteroaromatic nitro compounds. PTC includes 19 discrete labels and reports the carcinogenicity of 344 chemical compounds for male and female rats.

*Bioinformatic Datasets.* DD contains graphs of protein structures. A node represents an amino acid and edges are constructed if the distance of two nodes is less than 6. A label denotes whether a protein is an enzyme or non-enzyme. PROTEINS is a dataset where the nodes are secondary structure elements (SSEs), and there is an edge between two nodes if they are neighbors in the given amino acid sequence or in 3D space. The dataset has 3 discrete labels, representing helices, sheets or turns.

## B.2 Details of Molecular Datasets

*Input graph representation.* For simplicity, we use a minimal set of node and bond features that unambiguously describe the two-dimensional structure of molecules. We use RDKit [10] to obtain these features.

- Node features:
  - Atom number: [1, 118]
  - Chirality tag: {unspecified, tetrahedral cw, tetrahedral ccw, other}
- Edge features:
  - Bond type: {single, double, triple, aromatic}
  - Bond direction: {–, endupright, enddownright}

*Downstream task datasets.* 8 graph classification datasets from MoleculeNet [20] are used to evaluate model performance.

- BBBP [11]. Blood-brain barrier penetration (membrane permeability), involves records of whether a compound carries the permeability property of penetrating the blood-brain barrier.
- Tox21 [2]. Toxicity data on 12 biological targets, which has been used in the 2014 Tox21 Data Challenge and includes nuclear receptors and stress response pathways.
- ToxCast [15]. Toxicology measurements based on over 600 in vitro high-throughput screenings.

- SIDER [9]. Database of marketed drugs and adverse drug reactions (ADR), grouped into 27 system organ classes and also known as the Side Effect Resource.
- ClinTox [6, 13]. Qualitative data classifying drugs approved by the FDA and those that have failed clinical trials for toxicity reasons.
- MUV [5]. Subset of PubChem BioAssay by applying a refined nearest neighbor analysis, designed for validation of virtual screening techniques.
- HIV [1]. Experimentally measured abilities to inhibit HIV replication.
- BACE [18]. Qualitative binding results for a set of inhibitors of human  $\beta$ -secretase 1.

*Details of Dataset Splitting.* For molecular prediction tasks, following [14], we cluster molecules by scaffold (molecular graph substructure) [3], and recombine the clusters by placing the most common scaffolds in the training set, producing validation and test sets that contain structurally different molecules. Prior work has shown that this scaffold split provides a more realistic estimate of model performance in prospective evaluation compared to random split [4, 16]. The split for train/validation/test sets is 80%:10%:10%.

## C DETAILED EXPERIMENT SETUP

### C.1 Settings for Unsupervised Learning

**Datasets.** Eight benchmarks are adopted from TUDataset [12] and summarized in Table 1, including IMDB-BINARY, REDDIT-MULTI-5K, NCI1, MUTAG, PROTEINS, DD, REDDIT-BINARY, and COLLAB.

**Configuration.** Hidden dimension is chosen from {32, 64}, and batch size is chosen from {32, 128}. An Adam optimizer [8] is employed to minimize the contrastive lose with {0.01, 0.005, 0.001} learning rate.

**Learning protocols.** In unsupervised representation learning [19], all data is used for model pre-training and a non-linear SVM is adopted as classifier to perform to perform 10-fold cross-validation on learned graph embeddings. For graph representation learning, models are trained 20 epochs and tested every 10 epochs. The 10-fold evaluation are performed 5 times in total with different random

Table 2: Datasets statistics summary.

| Dataset | Category              | Utilization  | #Tasks | #Graphs   | Avg.Node | Avg.Degree |
|---------|-----------------------|--------------|--------|-----------|----------|------------|
| ZINC15  | Biochemical Molecules | Pre-Training |        | 2,000,000 | 26.63    | 57.72      |
| BBBP    | Biochemical Molecules | Finetuning   | 1      | 2,039     | 24.06    | 51.90      |
| Tox21   | Biochemical Molecules | Finetuning   | 12     | 7,831     | 18.57    | 38.58      |
| ToxCast | Biochemical Molecules | Finetuning   | 617    | 8,576     | 18.78    | 38.52      |
| SIDER   | Biochemical Molecules | Finetuning   | 27     | 1,427     | 33.64    | 70.71      |
| ClinTox | Biochemical Molecules | Finetuning   | 2      | 1,477     | 26.15    | 55.76      |
| MUV     | Biochemical Molecules | Finetuning   | 17     | 93,087    | 24.23    | 52.55      |
| HIV     | Biochemical Molecules | Finetuning   | 1      | 41,127    | 25.51    | 54.93      |
| BACE    | Biochemical Molecules | Finetuning   | 1      | 1,513     | 34.08    | 73.71      |

seeds as [19]. At last, we report the average accuracy and standard deviation (%).

## C.2 Setting for Transfer Learning

**Pre-training dataset.** ZINC15 [17] dataset is adopted for pre-training. In particular, a subset with two million unlabeled molecular graphs is sampled from the ZINC15.

**Pre-training details.** In the graph encoder setting in [7], GIN [21] with five convolutional layers is adopted for message passing. In particular, the hidden dimension is fixed to 300 across all layers and a pooling readout function that averages graph nodes is hired for NT-Xent loss calculation with the scale parameter  $\tau = 0.1$ . The hidden representations at the last layer are injected into the average pooling function. An Adam optimizer [8] is employed to minimize the integrated losses produced by the 5-layer GIN encoder. All training processes will run 100 epochs with a batch size of 256.

**Fine-tuning dataset.** We employ the eight ubiquitous benchmarks from the MoleculeNet dataset [20] as the downstream experiments. These benchmarks include a variety of molecular tasks like physical chemistry, quantum mechanics, physiology, and biophysics. For dataset split, the scaffold split scheme [4] is adopted for train/validation/test set generation. Table 2 summarizes the basic characteristics of the datasets.

**Fine-tuning details.** For downstream tasks, a linear layer is stacked after the pre-trained graph encoders for final property prediction. The downstream model still employs the Adam optimizer for 100 epochs of fine-tuning. All experiments on each dataset are performed for ten runs with different seeds, and the results are the averaged ROC-AUC scores (%)  $\pm$  standard deviations. The alternatives of learning rate in pre-training and fine-tuning phases are {0.0001, 0.001, 0.01}. To be in line with [22], the epochs for pre-training range from 20 to 100 with a step of 20 epochs.

## REFERENCES

- [1] [n. d.]. AIDS Antiviral Screen Data. ([n. d.]). Accessed: 2017-09-27, <https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data>.
- [2] 2014. Tox21 Data Challenge. (2014). Accessed: 2017-09-27, <https://tripod.nih.gov/tox21/challenge>.
- [3] Guy W Bemis and Mark A Murcko. 1996. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* 39, 15 (1996), 2887–2893.
- [4] Bin Chen, Robert P Sheridan, Viktor Hornak, and Johannes H Voigt. 2012. Comparison of random forest and Pipeline Pilot Naive Bayes in prospective QSAR predictions. *Journal of Chemical Information and Modeling* 52, 3 (2012), 792–803.
- [5] Eleanor J Gardiner, John D Holliday, Caroline O'Dowd, and Peter Willett. 2011. Effectiveness of 2D fingerprints for scaffold hopping. *Future Medicinal Chemistry* 3, 4 (2011), 405–414.
- [6] Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. 2016. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chemical Biology* 23, 10 (2016), 1294–1301.
- [7] W Hu, B Liu, J Gomes, M Zitnik, P Liang, V Pande, and J Leskovec. 2020. Strategies For Pre-training Graph Neural Networks. *International Conference on Learning Representations (ICLR)* (2020).
- [8] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.
- [9] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The SIDER database of drugs and side effects. *Nucleic Acids Research* 44, D1 (2016), D1075–D1079.
- [10] Greg Landrum. 2013. Rdkit documentation. *Release* 1, 1-79 (2013).
- [11] Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. 2012. A Bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of Chemical Information and Modeling* 52, 6 (2012), 1686–1697.
- [12] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. 2020. TUDataset: A collection of benchmark datasets for learning with graphs. *ICML 2020 Workshop on Graph Representation Learning and Beyond* (2020). arXiv:2007.08663
- [13] Paul A Novick, Oscar F Ortiz, Jared Poelman, Amir Y Abdulhay, and Vijay S Pande. 2013. SWEETLEAD: an in silico database of approved drugs, regulated chemicals, and herbal isolates for computer-aided drug discovery. *PLoS One* 8, 11 (2013), e79568.
- [14] Bharath Ramsundar, Peter Eastman, Patrick Walters, and Vijay Pande. 2019. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. O'Reilly Media.
- [15] Ann M Richard, Richard S Judson, Keith A Houck, Christopher M Grulke, Patra Volarath, Inthirany Thillainadarajah, Chihae Yang, James Rathman, Matthew T Martin, John F Wambaugh, et al. 2016. ToxCast chemical landscape: paving the road to 21st century toxicology. *Chemical Research in Toxicology* 29, 8 (2016), 1225–1251.
- [16] Robert P Sheridan. 2013. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling* (2013).
- [17] Teague Sterling and John J Irwin. 2015. ZINC 15—ligand discovery for everyone. *Journal of Chemical Information and Modeling* 55, 11 (2015), 2324–2337.
- [18] Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. 2016. Computational modeling of  $\beta$ -secretase 1 (BACE-1) inhibitors using ligand based approaches. *Journal of Chemical Information and Modeling* 56, 10 (2016), 1936–1949.
- [19] Fan-Yun Sun, Jordon Hoffman, Vikas Verma, and Jian Tang. 2020. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. *ICLR* (2020).
- [20] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* 9, 2 (2018), 513–530.
- [21] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks?. In *ICLR*.
- [22] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems* 33 (2020), 5812–5823.