Augment to Interpret: Supplementary Materials

1. Reusing a Module with Batch Normalisation

All our competitors use approximately the same GIN architecture for their experimentation (Suresh et al., 2021; Gao et al., 2022; Miao et al., 2022). Specifically, they all use batch normalisation layers, and so do we. However, considering the context it is used in, we argue that the way it should be used to be mathematically correct is not as straightforward as the way it is used in these approaches. This may have unexpected effects on their results and conclusions. Therefore, we propose a batch-norm switch to fix the issue, and we use it for our approach. As this is not within the scope of this paper, only a preliminary experiment is presented to demonstrate the relevance of a switch in practice. Further investigation into alternative methods to address the problem and the resulting effects on performances is left for future work.

1.1. Theoretical Analysis

In more detail, all mentioned approaches have one point in common: they use the same neural network to embed both the original graph (raw) and an augmented view of it (aug). However, the original graph and its augmented view come from different distributions, with significant differences. For example, the augmented view usually has lower node degrees, as only a subset of the original graph edges is kept. Ultimately, this difference has an impact on the input of the batch normalisation layers of this neural network.

As a reminder, the batch normalisation layer uses the batch mean and variance to normalise its input, and stores a running mean of these values to be used at test time as an estimate of said values. That is, the normalisation performed at train time is meant to be roughly equivalent to the normalisation performed at inference time. In that way, given an input, it produces roughly the same output, whether it is in train or in inference mode. This is important as the end of the network is trained on that output.

In our case, the mean and variance of the input of the layer are different whether the input of the network is a raw graph (μ_{raw}, σ_{raw}) or an augmented view of it (μ_{aug}, σ_{aug}). At train time, the normalisation will therefore be different in both cases. However, at our competitors' inference time, the estimate to perform the normalisation is something close to the average of the real values ($\frac{1}{2}(\mu_{raw} + \mu_{aug}), \frac{1}{2}(\sigma_{raw} + \sigma_{aug})$), for both the original graph and its augmented view. As a result, if the means and variances are very different, the outputs will not be similar in any way to the ones at train time. The final output of the network will therefore be highly unpredictable.

© 2023 .

1.2. Batch-Norm Switch and Preliminary Experiment

A way to avoid this issue is to have two different running means: one for the original graph distribution, and one for the augmented graph distribution. We call this solution a batch-norm switch, as it switches between running means. With a batch-norm switch, the training of the model remains the same as without the switch, but the inference is mathematically correct. To assess the relevance of the switch, a quick experiment is presented hereunder, based on the code provided by GSAT authors for the *BA2Motifs* dataset with a *GIN* architecture, using seed 0 only.

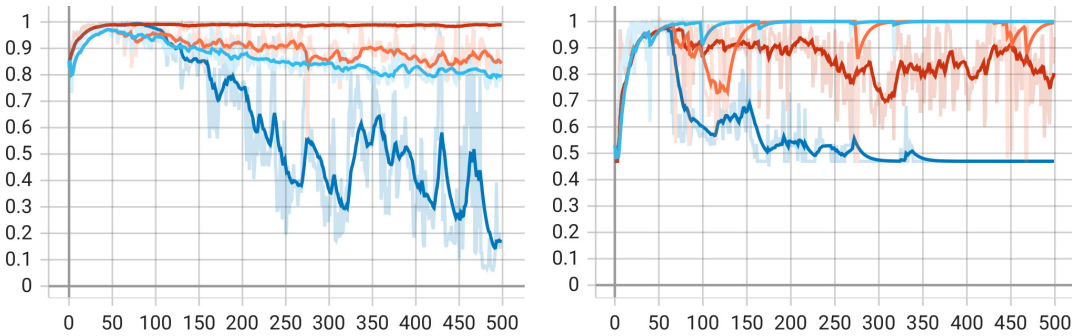
In GSAT, the issue is mitigated by having augmented views close to the original graphs, i.e. removing on average less than half the edges. It is further mitigated by an early stopping criterion based on a metric on the validation set, which is unlikely to give good results if the network produces unpredictable results. The early stopping will therefore tend to select epochs for which the issue has no negative impact.

For our analysis, we change these two parameters. First, we run the experiments for 500 epochs instead of 100 and show the full curves observed during training rather than the final result of the selected epoch. Second, for half the experiments, we use a final r value of 0.1 instead of 0.5. This means that starting from epoch 80, roughly 10% of edges are kept rather than 50%. Note the first 40 epochs are not affected by this second change, as GSAT adopt a step decay for r. The result is shown in Figure 1, measured on the test set.

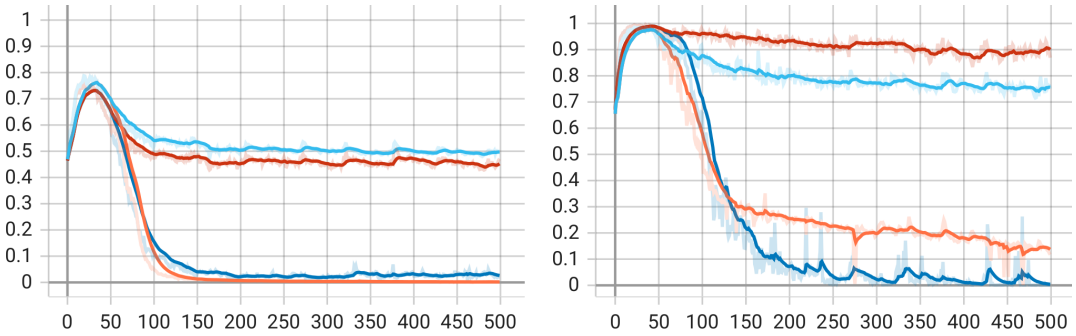
As can be seen, the downstream ACC is greatly improved by the switch. On the other hand, the interpretability AUC is improved by the switch when interpretations are sparse, but it is not as good when interpretations are not sparse. We think it is because the unexpected effect aligns with some dataset properties. For example, it could favour highly connected interpretations, which is good in the case of the *BA2Motifs* dataset but may not be for other datasets.

Additionally, we can see that the model with a switch and sparse interpretations (the orange curves) has on average almost null attention weights on non-important edges (0.0016 on average at the end), but significant weights on important edges (0.12 on average at the end). This indicates that the model misses important edges, but never includes non-important edges. For example, it could select length-3 cycles from class 1 and nothing from class 0, as shown in Figure 2. This could be considered a perfect interpretation, but it obtains a not-so-good interpretability AUC, given the expected interpretations include the length-5 cycles for each class.

A deeper analysis could be done to verify our hypotheses, challenge the batch-norm switch against alternatives, analyse results obtained with a batch-norm switch on each of our competitors or even challenge the *BA2Motifs* dataset itself, but this is out of the scope of this paper and is therefore left for future work.



(*a*) Interpretability AUC (*b*) Downstream ACC



(*c*) Avg. Background Attention Weights (*d*) Avg. Signal Attention Weights

Figure 1: Comparison of GSAT with a batch-norm switch (orange and teal) and without a batch-norm switch (red and dark blue) on *BA2Motifs*. The final value of r is set to 0.1 (dark blue and orange) or 0.5 (teal and red). A smoothing effect is applied to ease the analysis. Non-smoothed results are shown in transparent.

2. Beyond Topology-Based Interpretations

Our results indicate that there is a link between contrastive graph augmentation and interpretability. In such approaches, however, augmentations only focus on edges, and thus interpretations only highlight topology. A natural question is whether these conclusions stand for node features too. This section presents some early work on the matter.

We enhance the framework with an additional 2-layer MLP ν, trained to learn importance weights for node features. To make these weights understandable, we apply the feature augmentations in the input space, before being processed by the encoder.

The augmented feature matrix ($\widetilde{X}$) is obtained as $\widetilde{X} = (\nu(X)+\epsilon)\cdot X$ and with ϵ sampled as indicated in Section 3.1 with $\nu(X)$ a weight per node per feature, i.e. $\nu(X) \in \mathbb{R}^{N\times d}$.

We add an information loss on the importance scores to regularise the optimization, using the same r as the one for the topology information loss. As no ground truth is available for feature importance, we artificially add 20% of noisy features. We then average the importance score obtained by each feature on the test set and rank them. On *Mutag*, with our complete loss, we obtain the following results:

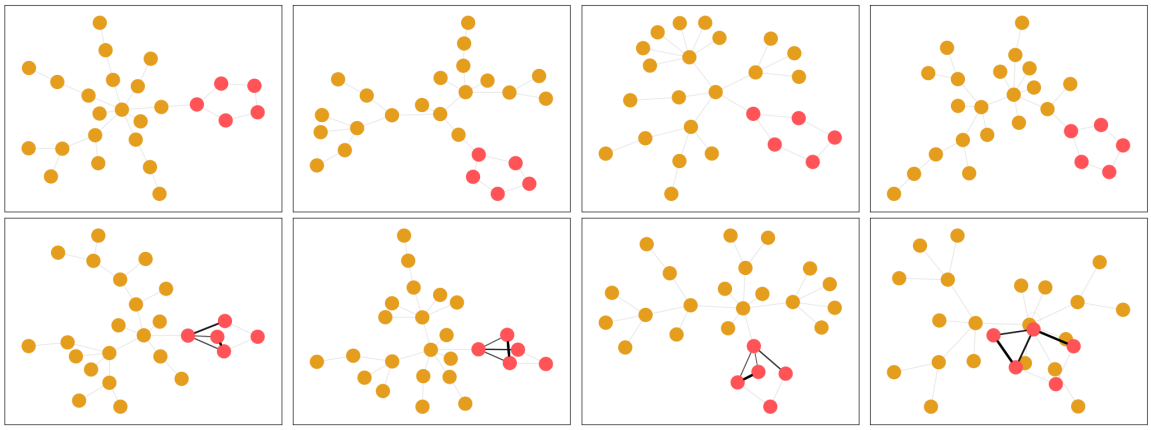


Figure 2: Interpretations of GSAT with a batch-norm switch and final r at 0.1 at epoch 370 for class 0 (top) and class 1 (bottom). Red nodes are part of the motif, and bold lines are predicted as important.

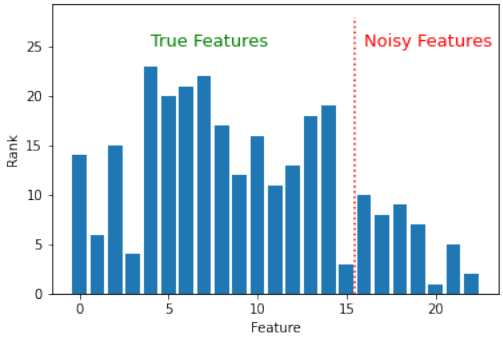
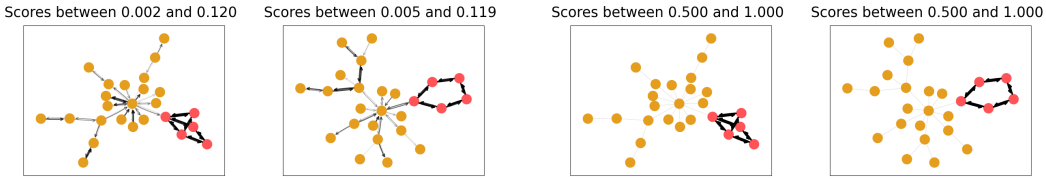


Figure 3: Ranking according to the average importance score given by the feature-augmentation module to each feature. It shows that noisy features are given less importance on average.

Figure 3 shows that the noisy features are given less importance on average, thus demonstrating an intrinsic feature selection ability which can be used as a feature importance scheme.

3. Impact of Sparsity on Human Readability

Figure 4(a) show some interpretations where important scores and unimportant ones are close. While truly important edges are highlighted, it is harder to distinguish them than in Figure 4(b)



(a) Sparse importance weights (b) Bi-modal importance weights

Figure 4: Examples of importance weights. A min-max normalisation is applied for clarity. Original weight limits are shown as the title of the figures.

4. Ablation Study of the Watchman Module

Our framework introduces a watchman module, with the purpose to stabilize the optimization. We observe in Table 3 that the presence of this module prevents quick drops in both interpretability AUC and downstream ACC, except for a few rare cases (as visible in Table 3 and Table 1).

This pattern is either positive or not significantly different across datasets, except for a few: INGENIOUS - Negative on *BA2Motifs* and certain methods on *SPMotifs.5* which significantly decrease due to conflicting optimization objectives, however we advise to always use the watchman for our regularised loss INGENIOUS ($\mathcal{L}$) as it improves the results on all datasets.

Interestingly, the loss which benefits the most from the watchman is the simple unregularised simclr loss ($\mathcal{L}$-info-negative), probably because in this case, the problem is underconstrained. It acts as a regularizer by enforcing the recovery of eigenvalues from the embeddings. This result shows the importance of regularisations of any kind for contrastive learning.

We conclude that while performance diminishes with overfitting as we just discussed, both the watchman and early stopping help alleviate this.

Moreover, defining the right number of training epochs is challenging in unsupervised settings. Our training curves (Figure 5) show that these unsupervised frameworks are prone to overfitting. This module can attenuate or delay this overfitting.

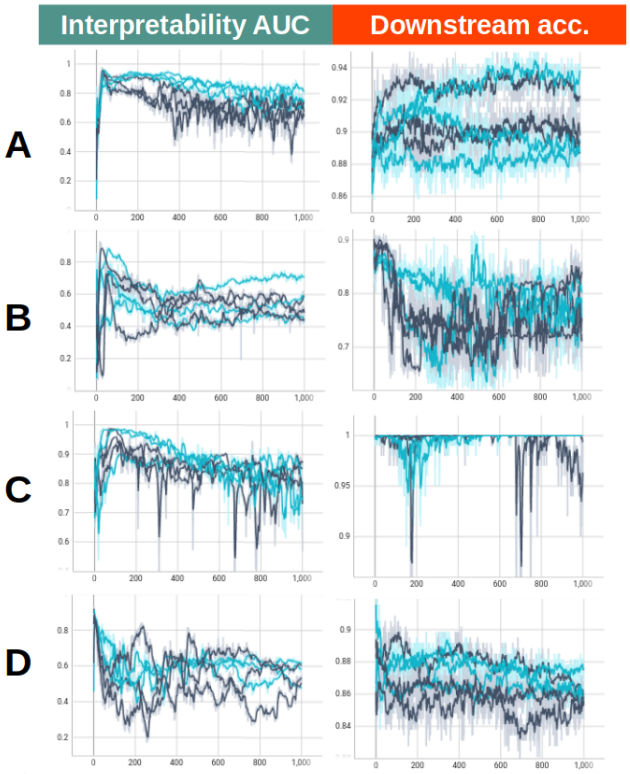


Figure 5: Example of training curves, giving interpretability AUC and downstream accuracy. Light blue is run with Watchman, dark blue without it. Those curves are: (A) for GSAT on *Mutag*, (B) for INGENIOUS on *Mutag*, (C) for INGENIOUS on *BA2Motifs*, (D) for the simple double augmentation simclr without regularisation on *Mutag*. The watchman helps prevent overtraining, on both supervised (GSAT) and unsupervised (INGENIOUS). The simple double-aug loss benefits too, showing that any form of regularisation is good.

5. Schematics and Other Figures

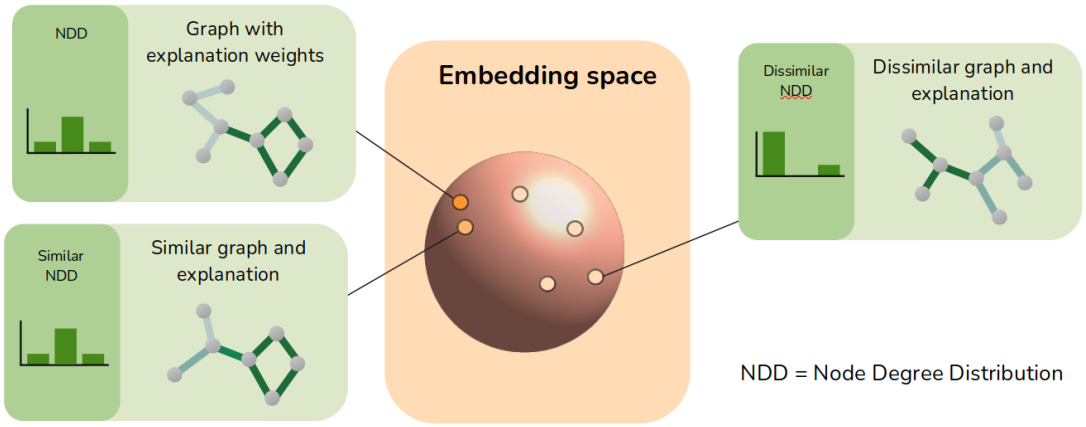


Figure 6: Principle of the Wasserstein distance metric. Our goal is to quantify the continuity of the interpretability functions. Therefore we evaluate whether similar graph embeddings obtain similar interpretations. Graph embedding proximity can be quantified by any vector similarity metric (we used Euclidian distance) and interpretation similarity can be quantified with any graph similarity metric (we use the Wasserstein distance between degree distributions).

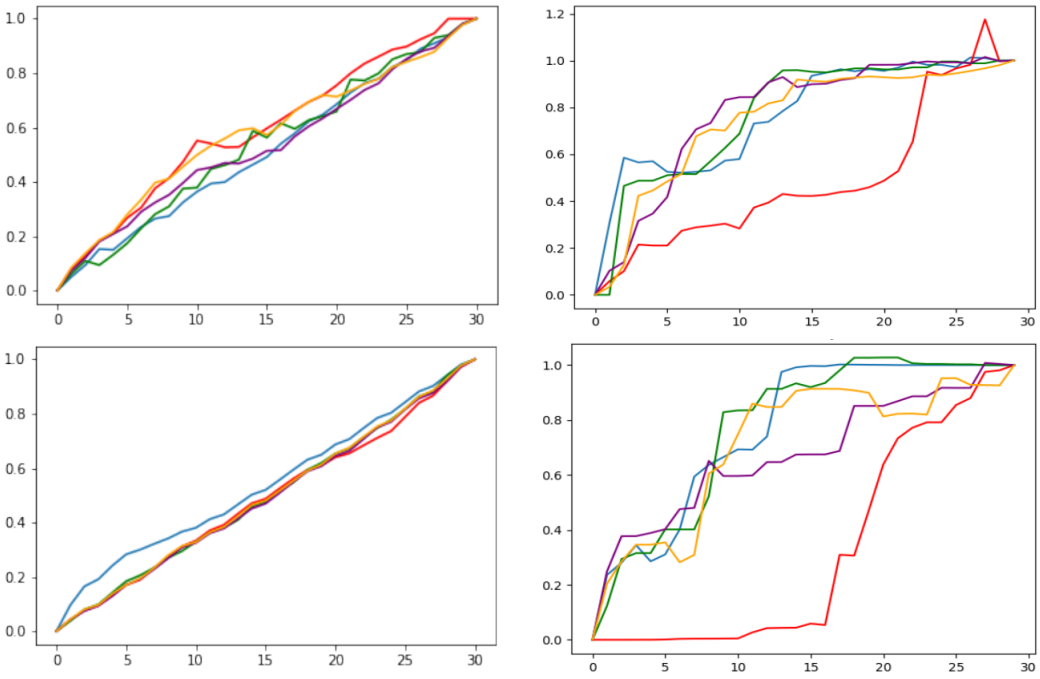


Figure 7: Examples of faithfulness curves on one graph of the *BA2Motifs* dataset. Clockwise starting from the top left: AD-GCL, GSAT, INGENIOUS, MEGA. We show the faithfulness (blue), faithfulness by removing edges according to ground truth (green), shuffled faithfulness (orange), random faithfulness (purple) and opposite faithfulness (red).

6. Full Tables

Wm.	Methods	Faithfulness Gap			Wasserstein Gap		
		BA2Motifs	Mutag	SPMotifs.5	BA2Motifs	Mutag	SPMotifs.5
	GSAT	$.34 \pm .00$	$.21 \pm .08$	$.06 \pm .02$	$.04 \pm .00$	$\underline{.02 \pm .00}$	$.04 \pm .00$
	AD-GCL	$.01 \pm .13$	$.03 \pm .09$	$.08 \pm .06$	$\underline{.10 \pm .01}$	$.01 \pm .00$	$.03 \pm .00$
	MEGA	$.06 \pm .02$	$.39 \pm .12$	$\underline{.68 \pm .03}$	$.04 \pm .01$	$.01 \pm .00$	$.01 \pm .00$
	INGENIOUS	$.39 \pm .05$	$.24 \pm .01$	$.50 \pm 0.00$	$.06 \pm .00$	$.01 \pm .00$	$\underline{.06 \pm 0.00}$
	INGENIOUS - Info	$.54 \pm .03$	$\underline{.58 \pm .04}$	$.46 \pm .07$	$.02 \pm .00$	$.01 \pm .00$	$.05 \pm .01$
	INGENIOUS - Negative	$.13 \pm .02$	$.19 \pm .02$	$.23 \pm .04$	$.02 \pm .00$	$.00 \pm .00$	$.02 \pm .00$
	INGENIOUS - Negative - Info	$.58 \pm .13$	$.17 \pm .13$	$.27 \pm .08$	$.04 \pm .03$	$.00 \pm .00$	$.02 \pm .00$
✓	GSAT	$.25 \pm .03$	$.06 \pm .10$	$-.03 \pm .01$	$.04 \pm .00$	$\underline{.02 \pm .00}$	$.04 \pm .00$
✓	INGENIOUS	$.60 \pm .02$	$.41 \pm .18$	$.45 \pm .03$	$.06 \pm .01$	$.01 \pm .00$	$.05 \pm .01$
✓	INGENIOUS - Info	$\underline{.68 \pm .02}$	$.48 \pm .03$	$.48 \pm .06$	$.05 \pm .02$	$.01 \pm .00$	$.05 \pm .01$
✓	INGENIOUS - Negative	$.13 \pm .02$	$.18 \pm .00$	$.22 \pm .06$	$.02 \pm .01$	$.01 \pm .00$	$.02 \pm .01$
✓	INGENIOUS - Negative - Info	$.61 \pm .10$	$.15 \pm .09$	$.32 \pm .00$	$.03 \pm .01$	$.01 \pm .00$	$.03 \pm .00$

Table 1: Table of full results. A check in the Wm column means the model uses the watchman loss. A higher faithfulness gap means the interpretations are more faithful, a higher wasserstein gap means the embedding space is more continuous in terms of interpretations.

Wm.	Loss	Downstream ACC			Intepretability AUC		
		Cora	Tree-cycle	Tree-grid	Cora	Tree-cycle	Tree-grid
	GSAT	$.610 \pm .07$	$.985 \pm .01$	$.984 \pm .01$	$.000 \pm .00$	$.630 \pm .17$	$.844 \pm .01$
	$\mathcal{L}$	$.572 \pm .06$	$.848 \pm .06$	$.954 \pm .02$	$.000 \pm .00$	$.232 \pm .11$	$.532 \pm .28$
	$\mathcal{L}$ - Info	$.593 \pm .05$	$.777 \pm .11$	$.952 \pm .03$	$.000 \pm .00$	$.256 \pm .09$	$.588 \pm .34$
	$\mathcal{L}$ - Negative	$.575 \pm .06$	$.951 \pm .01$	$.868 \pm .08$	$.000 \pm .00$	$.293 \pm .01$	$.549 \pm .04$
	$\mathcal{L}$ - Negative - Info	$.606 \pm .04$	$.856 \pm .02$	$.887 \pm .07$	$.000 \pm .00$	$.319 \pm .08$	$.702 \pm .07$

Table 2: Downstream ACC and interpretability AUC on node classification. Higher is better

Wm.	Loss	Downstream ACC			Intepretability AUC		
		BA2Motifs	Mutag	SPMotifs.5	BA2Motifs	Mutag	SPMotifs.5
	GSAT	$1 \pm .00$	$.921 \pm .02$	$.393 \pm .02$	$.998 \pm .00$	$.843 \pm .15$	$.895 \pm .01$
	AD-GCL	$1 \pm .00$	$.902 \pm .02$	$.337 \pm .00$	$.378 \pm .06$	$.416 \pm .19$	$.472 \pm .03$
	MEGA	$1 \pm .00$	$.886 \pm .01$	$.335 \pm .00$	$.459 \pm .28$	$.566 \pm .46$	$.506 \pm .03$
	$\mathcal{L}$	$1 \pm .00$	$.900 \pm .01$	$.366 \pm .01$	$.910 \pm .04$	$.745 \pm .11$	$.467 \pm .01$
	$\mathcal{L}$ - Info	$.990 \pm .00$	$.889 \pm .03$	$.361 \pm .02$	$.860 \pm .07$	$.606 \pm .23$	$.491 \pm .10$
	$\mathcal{L}$ - Negative	$1 \pm .00$	$.887 \pm .01$	$.340 \pm .02$	$.745 \pm .24$	$.546 \pm .04$	$.494 \pm .02$
	$\mathcal{L}$ - Negative - Info	$1 \pm .00$	$.895 \pm .01$	$.337 \pm .01$	$.462 \pm .38$	$.478 \pm .23$	$.494 \pm .04$
✓	GSAT	$1 \pm .00$	$.912 \pm .00$	$.389 \pm .01$	$.999 \pm .00$	$.904 \pm .04$	$.882 \pm .03$
✓	$\mathcal{L}$	$.993 \pm .01$	$.900 \pm .03$	$.363 \pm .00$	$.959 \pm .01$	$.771 \pm .16$	$.510 \pm .06$
✓	$\mathcal{L}$ - Info	$1 \pm .00$	$.863 \pm .04$	$.360 \pm .02$	$.900 \pm .02$	$.563 \pm .20$	$.499 \pm .07$
✓	$\mathcal{L}$ - Negative	$.995 \pm .01$	$.895 \pm .01$	$.338 \pm .00$	$.148 \pm .00$	$.620 \pm .03$	$.581 \pm .03$
✓	$\mathcal{L}$ - Negative - Info	$.997 \pm .01$	$.903 \pm .01$	$.337 \pm .00$	$.583 \pm .30$	$.650 \pm .18$	$.498 \pm .04$

Table 3: Downstream ACC and interpretability AUC on graph classification Higher is better.

Wm.	Loss	Random faithfulness			Sparsity		
		BA2Motifs	Mutag	SPMotifs.5	BA2Motifs	Mutag	SPMotifs.5
	GSAT	$.840 \pm .04$	$.731 \pm .03$	$.708 \pm .07$	$.883 \pm .01$	$.812 \pm .14$	$.874 \pm .00$
	AD-GCL	$.606 \pm .02$	$.774 \pm .03$	$.679 \pm .01$	$.608 \pm .08$	$.955 \pm .05$	$1 \pm .00$
	MEGA	$.486 \pm .01$	$.638 \pm .02$	$.614 \pm .01$	$.566 \pm .13$	$.549 \pm .07$	$.275 \pm .05$
	$\mathcal{L}$	$.875 \pm .02$	$.804 \pm .03$	$.649 \pm .01$	$.523 \pm .03$	$.897 \pm .11$	$.180 \pm .01$
	$\mathcal{L}$ - Info	$.714 \pm .09$	$.666 \pm .08$	$.651 \pm .03$	$.266 \pm .12$	$.191 \pm .20$	$.166 \pm .01$
	$\mathcal{L}$ - Negative	$1.020 \pm .04$	$.859 \pm .01$	$.647 \pm .06$	$.991 \pm .00$	$.980 \pm .00$	$.714 \pm .24$
	$\mathcal{L}$ - Negative - Info	$.897 \pm .13$	$.879 \pm .02$	$.641 \pm .06$	$.417 \pm .13$	$.958 \pm .07$	$.657 \pm .29$
✓	GSAT	$.817 \pm .02$	$.822 \pm .01$	$.538 \pm .05$	$.855 \pm .04$	$.893 \pm .00$	$.848 \pm .00$
✓	$\mathcal{L}$	$.719 \pm .01$	$.783 \pm .02$	$.636 \pm .01$	$.402 \pm .04$	$.591 \pm .31$	$.205 \pm .06$
✓	$\mathcal{L}$ - Info	$.743 \pm .03$	$.670 \pm .09$	$.606 \pm .02$	$.288 \pm .04$	$.269 \pm .23$	$.168 \pm .01$
✓	$\mathcal{L}$ - Negative	$.877 \pm .03$	$.808 \pm .02$	$.663 \pm .07$	$.840 \pm .02$	$.969 \pm .00$	$.808 \pm .31$
✓	$\mathcal{L}$ - Negative - Info	$.767 \pm .10$	$.789 \pm .05$	$.594 \pm .02$	$.402 \pm .09$	$.971 \pm .04$	$.466 \pm .10$

Table 4: Random faithfulness and sparsity on graph classification. For random faithfulness, higher is better and for sparsity lower is better.

Wm.	Loss	faithfulness			Opposite faithfulness		
		Cora	Tree-cycle	Tree-grid	Cora	Tree-cycle	Tree-grid
	GSAT	$.756 \pm .01$	$.728 \pm .21$	$.960 \pm .03$	$.550 \pm .07$	$.761 \pm .05$	$.691 \pm .01$
	$\mathcal{L}$	$.747 \pm .03$	$.839 \pm .07$	$.965 \pm .07$	$.628 \pm .13$	$.699 \pm .08$	$.707 \pm .05$
	$\mathcal{L}$ - Info	$.752 \pm .03$	$.767 \pm .03$	$.868 \pm .06$	$.611 \pm .10$	$.680 \pm .17$	$.584 \pm .11$
	$\mathcal{L}$ - Negative	$.732 \pm .01$	$.853 \pm .08$	$.902 \pm .03$	$.531 \pm .06$	$.846 \pm .05$	$.610 \pm .15$
	$\mathcal{L}$ - Negative - Info	$.783 \pm .04$	$.780 \pm .04$	$.847 \pm .01$	$.477 \pm .03$	$.608 \pm .05$	$.598 \pm .11$

Table 5: Faithfulness and opposite faithfulness on node classification. For faithfulness, higher is better and for opposite faithfulness lower is better.

Wm.	Loss	Random faithfulness			Sparsity		
		Cora	Tree-cycle	Tree-grid	Cora	Tree-cycle	Tree-grid
	GSAT	$.693 \pm .06$	$.814 \pm .12$	$.911 \pm .02$	$.708 \pm .01$	$.827 \pm .04$	$.801 \pm .00$
	$\mathcal{L}$	$.727 \pm .02$	$.823 \pm .13$	$.877 \pm .04$	$.754 \pm .06$	$.629 \pm .08$	$.639 \pm .06$
	$\mathcal{L}$ - Info	$.727 \pm .03$	$.757 \pm .07$	$.816 \pm .10$	$.711 \pm .04$	$.481 \pm .03$	$.565 \pm .11$
	$\mathcal{L}$ - Negative	$.718 \pm .03$	$.818 \pm .08$	$.849 \pm .01$	$.669 \pm .04$	$.827 \pm .01$	$.609 \pm .12$
	$\mathcal{L}$ - Negative - Info	$.714 \pm .04$	$.735 \pm .06$	$.816 \pm .03$	$.740 \pm .01$	$.564 \pm .10$	$.592 \pm .05$

Table 6: Random Faithfulness and Sparsity on node classification. For random faithfulness, higher is better and for sparsity lower is better.

Wm.	Method	W_1 global			W_1 local		
		BA2Motifs	Mutag	SPMotifs.5	BA2Motifs	Mutag	SPMotifs.5
	GSAT	$.065 \pm .02$	$.097 \pm .00$	$.080 \pm .00$	$.019 \pm .01$	$.071 \pm .01$	$.036 \pm .00$
	AD-GCL	$.240 \pm .04$	$.099 \pm .00$	$.117 \pm .03$	$.134 \pm .02$	$.085 \pm .00$	$.082 \pm .02$
	MEGA	$.139 \pm .04$	$.128 \pm .01$	$.115 \pm .00$	$.095 \pm .03$	$.113 \pm .01$	$.099 \pm .01$
	INGENIOUS	$.154 \pm .01$	$.099 \pm .00$	$.113 \pm .01$	$.093 \pm .02$	$.087 \pm .01$	$.044 \pm .01$
	INGENIOUS - Info	$.144 \pm .01$	$.096 \pm .01$	$.105 \pm .02$	$.122 \pm .01$	$.084 \pm .00$	$.047 \pm .01$
	INGENIOUS - Negative	$.123 \pm .01$	$.100 \pm .00$	$.132 \pm .01$	$.101 \pm .01$	$.095 \pm .00$	$.103 \pm .01$
	INGENIOUS - Negative - Info	$.144 \pm .02$	$.100 \pm .00$	$.117 \pm .01$	$.098 \pm .02$	$.092 \pm .00$	$.092 \pm .02$
✓	GSAT	$.055 \pm .01$	$.097 \pm .00$	$.080 \pm .00$	$.012 \pm .00$	$.077 \pm .00$	$.035 \pm .00$
✓	INGENIOUS	$.143 \pm .01$	$.099 \pm .00$	$.095 \pm .01$	$.080 \pm .01$	$.084 \pm .00$	$.044 \pm .00$
✓	INGENIOUS - Info	$.151 \pm .01$	$.100 \pm .00$	$.094 \pm .02$	$.098 \pm .01$	$.087 \pm .00$	$.042 \pm .00$
✓	INGENIOUS - Negative	$.136 \pm .00$	$.101 \pm .00$	$.116 \pm .01$	$.107 \pm .02$	$.089 \pm .00$	$.088 \pm .02$
✓	INGENIOUS - Negative - Info	$.140 \pm .00$	$.101 \pm .00$	$.110 \pm .01$	$.105 \pm .01$	$.087 \pm .00$	$.074 \pm .01$

Table 7: Global and local Wasserstein distances on graph embeddings.

Wm.	Method	Faithfulness			Opposite Faithfulness		
		BA2Motifs	Mutag	SPMotifs.5	BA2Motifs	Mutag	SPMotifs.5
	GSAT	$.843 \pm .02$	$.778 \pm .05$	$.702 \pm .06$	$.497 \pm .02$	$.565 \pm .04$	$.643 \pm .03$
	AD-GCL	$.622 \pm .07$	$.779 \pm .00$	$.707 \pm .02$	$.609 \pm .06$	$.747 \pm .09$	$.626 \pm .05$
	MEGA	$.531 \pm .01$	$.827 \pm .04$	$.898 \pm .01$	$.467 \pm .02$	$.436 \pm .09$	$.214 \pm .02$
	INGENIOUS	$.831 \pm .03$	$.830 \pm .03$	$.832 \pm 0.01$	$.432 \pm .05$	$.588 \pm .04$	$.325 \pm 0.01$
	INGENIOUS - Info	$.823 \pm .01$	$.864 \pm .03$	$.817 \pm .06$	$.274 \pm .04$	$.279 \pm .03$	$.356 \pm .03$
	INGENIOUS - Negative	$.978 \pm .05$	$.834 \pm .02$	$.744 \pm .03$	$.844 \pm .03$	$.637 \pm .00$	$.506 \pm .07$
	INGENIOUS - Negative - Info	$.912 \pm .04$	$.838 \pm .04$	$.746 \pm .01$	$.323 \pm .12$	$.664 \pm .10$	$.474 \pm .10$
✓	GSAT	$.782 \pm .02$	$.747 \pm .07$	$.511 \pm .04$	$.523 \pm .02$	$.684 \pm .04$	$.548 \pm .05$
✓	INGENIOUS	$.856 \pm .01$	$.862 \pm .07$	$.797 \pm .02$	$.257 \pm .01$	$.443 \pm .12$	$.345 \pm .01$
✓	INGENIOUS - Info	$.884 \pm .02$	$.819 \pm .07$	$.796 \pm .02$	$.200 \pm .01$	$.336 \pm .03$	$.316 \pm .04$
✓	INGENIOUS - Negative	$.801 \pm .00$	$.803 \pm .01$	$.738 \pm .02$	$.663 \pm .02$	$.617 \pm .02$	$.518 \pm .08$
✓	INGENIOUS - Negative - Info	$.893 \pm .05$	$.782 \pm .03$	$.730 \pm .01$	$.279 \pm .08$	$.629 \pm .08$	$.409 \pm .02$

Table 8: Faithfulness depending on experiments. For faithfulness, higher is better, for opposite faithfulness, lower is better.

References

Hang Gao, Jiangmeng Li, Wenwen Qiang, Lingyu Si, Fuchun Sun, and Changwen Zheng. Bootstrapping informative graph augmentation via a meta learning approach. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3001–3007. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. *International Conference on Machine Learning*, 2022.

Susheel Suresh, Pan Li, Cong Hao, and Jennifer Neville. Adversarial graph augmentation to improve graph contrastive learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.