Additional file 1:
Supplementary notes and summary tables

MolSanity: A Reliability Audit for Feature Attributions on Molecular Graph Neural Networks

Every table in this file is generated from the committed artifacts of the single run reported in the manuscript, by `paper/make_additional_files.py`. Additional files 2–4 carry the full per-cell data in machine-readable form.

Contents

S1 Ground-truth benchmark by dataset and split **2**

S2 Does the model read the ground truth? **3**

S3 When not to trust an attribution **3**

S4 Machine-readable additional files **4**

S1 Ground-truth benchmark by dataset and split

Computed from per-molecule audit records under `artifacts/audit/`; every number is computed, none fixed. **Question:** does ranking attributors by a faithfulness / fidelity metric — what SOTA evaluation frameworks (GraphFramEx, MolFaith, DIG) emit — recover the attributor the *ground truth* says is best? We contrast two regimes: in-distribution vs shift.

`occ_spearman` = MolSanity occlusion faithfulness · `Fidelity+` and `characterization` = field-standard / GraphFramEx · `rank corr` ρ = Spearman correlation between the faithfulness metric and GT AUROC across attributors ($\approx 1 \rightarrow$ faithfulness tracks correctness; $\leq 0 \rightarrow$ it does not).

MUTAG · GINE · random split — in-distribution (motif-proxy GT)

7 attributors on the same ~58 molecules.

attributor	GT AUROC	95% CI	occ_spearman	Fidelity+	characterization
GNNExplainer ⋆	0.858	(0.791, 0.921)	-0.170	0.310	0.338
IntegratedGradients	0.496	(0.469, 0.523)	-0.196	0.551	0.347
SubgraphX	0.348	(0.316, 0.378)	-0.154	0.342	0.450
PGExplainer	0.251	(0.223, 0.278)	-0.351	0.379	0.225
GuidedBackprop	0.037	(0.018, 0.066)	-0.089	0.540	0.178
InputXGradient	0.013	(0.005, 0.023)	-0.200	0.491	0.160
Saliency	0.002	(0.000, 0.005)	-0.195	0.494	0.163

⋆ = attributor the exact/proxy ground truth ranks best.

Faithfulness-only selection test — would a framework ranking by each metric pick the GT-best attributor?

faithfulness metric	its top pick	pick GT AUROC	GT-best	GT-best AUROC	mismatch?	paired Wilcoxon p	rank corr ρ(faith,GT)
occ_spearman	GuidedBackprop	0.037	GNNExplainer	0.858	**yes**	0.0000	0.143
fidelity_plus	IntegratedGradients	0.496	GNNExplainer	0.858	**yes**	0.0000	-0.357
characterization	SubgraphX	0.348	GNNExplainer	0.858	**yes**	0.0000	0.821

MUTAG · GINE · scaffold split — scaffold shift (motif-proxy GT)

7 attributors on the same ~53 molecules.

attributor	GT AUROC	95% CI	occ_spearman	Fidelity+	characterization
GNNExplainer ⋆	0.826	(0.754, 0.896)	0.379	0.154	0.209
PGExplainer	0.360	(0.298, 0.429)	0.281	0.066	0.114
SubgraphX	0.330	(0.294, 0.366)	0.379	0.304	0.426
IntegratedGradients	0.302	(0.235, 0.367)	0.310	0.030	0.077
InputXGradient	0.079	(0.060, 0.099)	0.499	0.167	0.206
GuidedBackprop	0.013	(0.007, 0.019)	0.616	0.256	0.323
Saliency	0.002	(0.000, 0.003)	0.404	0.157	0.202

⋆ = attributor the exact/proxy ground truth ranks best.

Faithfulness-only selection test — would a framework ranking by each metric pick the GT-best attributor?

faithfulness metric	its top pick	pick GT AUROC	GT-best	GT-best AUROC	mismatch?	paired Wilcoxon p	rank corr ρ(faith,GT)
occ_spearman	GuidedBackprop	0.013	GNNExplainer	0.826	**yes**	0.0000	-0.643
fidelity_plus	SubgraphX	0.330	GNNExplainer	0.826	**yes**	0.0000	-0.321
characterization	SubgraphX	0.330	GNNExplainer	0.826	**yes**	0.0000	0.036

What this shows

- **MUTAG · GINE · random split**: ranking by faithfulness picks the wrong attributor on 3 of 3 metrics (ρ ranges -0.36 to +0.82).
- **MUTAG · GINE · scaffold split**: ranking by faithfulness picks the wrong attributor on 3 of 3 metrics (ρ ranges -0.64 to +0.04).

Mean rank correlation falls from +0.20 in-distribution to -0.31 under scaffold shift (-0.51). Faithfulness and correctness dissociate under shift, so a faithfulness-only benchmark can recommend the wrong method in exactly the regime that matters for drug discovery.

Single-seed figures. Read `SEED_VARIANCE.md` before treating any attributor ranking above as an effect.

S2 Does the model read the ground truth?

Faber et al. (KDD 2021) argue that scoring attributions against a known rationale misleads when the trained model did not use that rationale: a low GT AUROC would then be a fact about the model, not the explanation. This is testable. Occlude the ground-truth substructure; if the prediction collapses, the model *is* using it.

- molecules where the model reads the ground truth: **7524**
- molecules where it does not (Faber applies): **8085**
- mean GT AUROC when the model reads it: **0.781**
- mean GT AUROC when it does not: **0.719**

The number that answers the objection

1395 molecules (0.185 of those the model demonstrably reads the ground truth from) still receive an attribution anti-aligned with it. On those, no appeal to an alternative rationale explains the result: the attribution misdescribes a model that is provably using the substructure the attribution ranks lowest.

S3 When not to trust an attribution

The audit cannot tell a practitioner which attributor to pick without ground truth. It can tell them when to decline to trust one. Each signal below is available at inference time; molecules with the worst values are abstained on, and the table reports the ground-truth localisation of what remains.

`lift` is the gain in mean GT AUROC between keeping everything and keeping the best half. **A signal with lift <= 0 is useless for abstention on this data**, which is itself worth knowing: it means the corresponding intuition is unsupported.

signal	n	GT AUROC @100%	@50%	lift	below chance @100%	@50%
occ_spearman	15201	0.748	0.804	+0.056	0.216	0.151
confidence	15609	0.749	0.793	+0.044	0.214	0.168
rationale_reliance	15609	0.749	0.782	+0.032	0.214	0.184
stability	14316	0.764	0.784	+0.021	0.191	0.166
motif_top1_share	15609	0.749	0.667	-0.082	0.214	0.323

Recommended rule

Rank molecules by **occ_spearman** and abstain on the tail. No coverage level gets the below-chance share under 10%. The best achievable is **14.5%** at 37% coverage. Abstention narrows the problem here but does not solve it, and that is the honest reading. Mean GT AUROC already exceeds 0.70 at full coverage (0.748), so a rule stated against the mean would be vacuous. The pooled mean hides the tail; the below-chance share does not.

The transfer assumption. These curves are computed where ground truth exists. Applying the rule to a real molecular dataset assumes the signal-reliability relationship carries over to cells where correctness cannot be measured. This paper's central finding is that such transfer fails across splits, so the assumption is stated here rather than relied on silently.

S4 Machine-readable additional files

- **Additional file 2** (`Additional_file_2.csv`) — Per-cell audit matrix One row per (dataset, backbone, attributor, split, seed) combination: ground-truth localisation and AUPRC where node labels exist, occlusion faithfulness, motif coherence, Fidelity+/-, sparsity, calibration error and the predictive metrics. This is the table every figure and every quoted statistic in the manuscript is computed from. *Columns:* `section`, `dataset`, `backbone`, `attributor`, `split`, `seed`, `n_mol`, `acc`, `auc`, `gt_auroc`, `gt_auprc`, `motif_top1`, `occ_spearman`, `occ_top1`, `fid+`, `fid-`, `sparsity`, `ece`, `rmse`, `mae`, `r2`.
- **Additional file 3** (`Additional_file_3.csv`) — Across-seed variance per cell Mean, standard deviation, minimum and maximum over the three seeds for every cell run under more than one seed. The spread here is the check on the single-seed selection tests of Table 5. *Columns:* `section`, `dataset`, `backbone`, `attributor`, `split`, `mean`, `sd`, `min`, `max`.
- **Additional file 4** (`Additional_file_4.csv`) — Head-to-head benchmark including framework metrics Every audited cell scored on MolSanity's axes alongside the field-standard Fidelity+/-, sparsity, the GraphFramEx characterisation score and the PyG/DIG unfaithfulness metric, computed on the same molecules. *Columns:* `section`, `dataset`, `backbone`, `attributor`, `split`, `n_mol`, `gt_auroc`, `occ_spearman`, `stability`, `motif_top1_share`, `fidelity_plus`, `fidelity_minus`, `sparsity`, `characterization`, `unfaithfulness`, `method A`, `method B`, `n`, `median Δ(A-B)`, `p-value`.

Each file carries the union of the columns of the tables inside it, with `section` naming the table a row came from; a cell is empty where the metric does not exist for that row, rather than carrying a dash, so the files parse directly as data.