Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Karthik Iyer

22557016@students.latrobe.edu.au

La Trobe University

Nasser Sabar

La Trobe University

Method Article

Keywords:

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10621100/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

Karthik Iyer[1,*] Nasser Sabar[1]

[1]Department of Computer Science and Information Technology, La Trobe University, Melbourne, Victoria 3086, Australia

*Correspondence: 22557016@students.latrobe.edu.au; karthik.iyer@hotmail.com ORCID: 0009-0004-0593-1602

Abstract. Feature attributions are routinely used to justify molecular graph neural network (GNN) predictions to chemists, yet they are almost never audited for *reliability*: existing frameworks ask whether an explanation is faithful to the model, not whether it identifies the chemistry that determines the property, nor where it stops being trustworthy under scaffold shift. **MolSanity** is a reliability-audit framework that wraps canonical implementations (Captum, PyTorch Geometric, RDKit) rather than proposing a new attributor, scoring every (dataset $\times$ backbone $\times$ attributor $\times$ split) cell on six axes: motif-native coherence, occlusion–attribution faithfulness, ground-truth localisation where node labels exist, cross-checkpoint stability, calibration linkage, and confidence/correctness regime stratification. Our central finding is that *faithfulness is not correctness*: across this matrix the two show no dependable relationship in either regime. In 30 selection tests — 5 molecular ground-truth arms $\times$ two splits $\times$ 3 ranking metrics — a faithfulness-only ranking picks an attributor other than the ground-truth-best one in 26, no less often in distribution (14 of 15) than under shift (12 of 15); on MUTAG under shift it prefers an attributor anti-aligned with the nitro motif (GT AUROC 0.013) over one at 0.826. Pooled over 47 cells the faithfulness–correctness rank correlation is $+0.222$ in distribution ($p = 0.134$) and -0.124 under shift ($p = 0.405$), neither distinguishable from zero, while per arm under shift it runs from -0.564 to $+0.786$ — a spread 5 arms cannot separate from noise ($Q = 8.53$, $p = 0.074$). Either way no pooled figure survives leave-one-arm-out, so a reliability claim must be indexed to the individual (dataset, backbone, attributor, split) cell. Restricted to the 3 arms of an earlier analysis it gives -0.356 ($p = 0.042$); two externally authored rationale benchmarks remove the effect, and we report the 5-arm result. Faithfulness itself does not fall under shift — it *rises*, from 0.049 to 0.132, while localisation does not move — so the metric gives no warning. Over 31785 per-molecule records, localisation degrades on confidently-wrong predictions (0.769 to 0.681) while their faithfulness *improves*, and the calibration–reliability link attenuates from a per-cell median of 0.144 to 0.074 when pooled. Every number, figure and table regenerates from the committed artifacts.

1 Introduction

A medicinal chemist is shown a graph neural network's prediction that a candidate is hERG-active, together with a heat map over its atoms. The heat map concentrates on a piperidine nitrogen; the chemist finds that plausible, and the series is deprioritised. Nothing in that workflow ever established that the highlighted atoms are the atoms the model actually used, that they are the atoms that matter chemically, or that the highlighting behaves the same way on the next scaffold. Each of those is a separate question, and the answers come apart.

They come apart in a way that is measurable, and that gets worse in exactly the setting drug discovery operates in. An attribution can be *faithful*, in that the atoms it ranks highest really are the atoms whose removal moves the model's output most, and simultaneously *wrong*, in that those atoms are not the substructure that determines the property. Across the 142 committed cells in this study that carry a node-level ground truth, 14 are positively faithful to the model and yet *anti*-aligned with the true motif (ground-truth AUROC below the chance value of 0.5). A benchmark that scores only faithfulness rates all of them as fine.

In practice the problem takes the form of a selection problem. With no ground truth available, which is the situation on every real molecular dataset in this study, an attributor has to be chosen using a faithfulness or fidelity metric. On the 5 molecular cell families that do have a node-level ground truth and were audited on both splits, we can check whether that choice is the right one. It usually is not: across 30 selection tests — 5 arms $\times$ two splits $\times$ 3 ranking metrics — the metric's top pick differs from the ground-truth-best attributor in 26. The worst case is not subtle. Under a Bemis–Murcko scaffold split on MUTAG, all 3 ranking metrics miss: our occlusion measure selects GuidedBP, whose agreement with the mutagenic nitro motif is GT AUROC 0.013, while Fidelity+ and the GraphFramEx characterisation score both select SubgraphX at 0.330 — against GNNExpl., which the ground truth ranks best at 0.826. The median per-molecule gap is 0.967 AUROC (paired Wilcoxon $p < 0.001$, $n = 53$; Benjamini–Hochberg $q < 0.001$ across the 30 tests).

Crucially, this is not a distribution-shift artefact. The mismatch rate is 14 of 15 in distribution against 12 of 15 under scaffold shift, so an in-distribution benchmark offers no protection.

None of this makes faithfulness metrics wrong. They answer a different question from the one a chemist is asking. What we can say about the relationship between the two is mostly negative: pooled over 47 molecular ground-truth cells the faithfulness–correctness rank correlation is $+0.222$ in distribution ($p = 0.134$) and -0.124 under scaffold shift ($p = 0.405$), neither distinguishable from zero, while the per-arm coefficients under shift run from -0.564 to $+0.786$ under an identical protocol. Reliability is a property of the cell — the (dataset, backbone, attributor, split) combination — and not of the attributor, and a single headline number about attribution quality is the wrong shape of claim. Molecular property prediction is nonetheless deployed under scaffold shift almost by construction, since the point of a screening model is to score series the training set does not contain, and scaffold-split evaluation is now standard practice for the *predictions* themselves [26, 27]. The evaluation protocol for the *explanations* has not followed: it does not stratify by that shift at all, which is why we can report the stratification as a finding rather than assume it.

Contributions. This paper presents **MolSanity**, a reliability-audit framework for feature attributions on molecular GNNs. Rather than propose another attribution method, it wraps the canonical implementations (Captum [17], PyTorch Geometric [18], RDKit [19]) and audits them. Specifically:

- We define a six-axis audit over a motif-native (RDKit) decomposition: coherence, occlusion–attribution faithfulness, ground-truth localisation, cross-checkpoint stability, calibration linkage and confidence/correctness regime stratification, with field-standard Fidelity±, sparsity, the GraphFramEx characterisation score [11] and the PyG/DIG unfaithfulness metric [12] computed on the *same* molecules for direct comparability (The MolSanity audit).
- We show that a faithfulness-only ranking picks the wrong attributor in 26 of 30 tests spanning 5 molecular ground-truth arms and both splits, in a design where within an arm only the split varies, so the regime comparison is not confounded with a change of dataset. The failure is not confined to shift: it is 14 of 15 in distribution (A fidelity-only ranking picks the wrong attributor, in both regimes).
- We separate what shift breaks from what it leaves alone, and on this run it leaves a great deal alone. Which *backbone* localises best is largely preserved across the split ($\rho = 0.700$–0.900 over the 3 molecular arms carrying all 5 backbones), though *which* backbone is used moves localisation by 0.125–0.538 AUROC and so has to be reported. The *attributor* ordering largely survives too ($\rho = 0.821$ on MUTAG, 0.964 on MolMotif), and the level of faithfulness does not shift systematically either (across the 93 cells audited on both splits the median change is 0.077). The *relationship* between faithfulness and correctness — the quantity a faithfulness-only benchmark cannot see — is absent in both regimes rather than broken by one of them, and varies by arm from -0.564 to $+0.786$ under shift.
- Using 31785 per-molecule audit records, we stratify reliability by confidence/correctness regime and test the calibration-linkage hypothesis directly. It holds per cell (median $\rho = 0.144$, 85 cells significantly positive vs. 29 negative) but *attenuates to* 0.074 when cells are pooled, a Simpson's-paradox trap for anyone reporting a single aggregate number.
- We report negative as well as positive findings. Adding two externally authored molecular rationale benchmarks changes the pooled shift correlation from -0.356 ($p = 0.042$) over 3 arms to -0.124 ($p = 0.405$) over 5. Also here: an occlusion operator that is mis-specified on regression in a way that mimics a finding; non-molecular ground-truth arms whose scaffold split is degenerate and are therefore excluded from the shift contrast rather than counted toward it; an exactly-labelled molecular arm too saturated to adjudicate anything on its own; and the 44 cells where the correctness axis does not exist and is marked as such rather than scored.

2 Related work

Attribution methods. The attributors audited here are the standard ones. The gradient family (Saliency [2], Input×Gradient, Guided Backpropagation [3] and Integrated Gradients [1]) is model-agnostic and cheap. Graph-specific methods learn a mask over the input: GNNExplainer [4] optimises a soft edge/feature mask per instance, PGExplainer [5] amortises that into a parametric mask network, and SubgraphX [6] searches connected subgraphs with Monte-Carlo tree search under a Shapley objective. Graph analogues of CAM/Grad-CAM were introduced by Pope et al. [7]; Yuan et al. [8] survey the space. In every case we call the published implementation rather than reimplementing it.

Evaluating explanations. Two evaluation traditions exist. *Faithfulness* measures whether the explanation tracks

the model: Fidelity± and sparsity [8], the GraphFramEx characterisation score that combines them [11], and the unfaithfulness metric shipped with DIG/PyG [12]. *Correctness* measures whether the explanation recovers a known rationale, which requires ground-truth substructures, typically synthetic [10]. The tension between the two is known: Faber et al. [14] argue that comparing against ground truth is misleading when the trained model does not use that rationale, and Sanchez-Lengeling et al. [9] built the first systematic attribution benchmark for molecular GNNs on exactly this basis. Outside graphs, sanity checks [15] and retraining-based benchmarks [16] made a parallel case that plausible-looking attributions can be uninformative. MolSanity's position is that neither axis subsumes the other, and that what needs measuring is the *joint* behaviour of the two under distribution shift.

Frameworks, and the explicit delta. Four lines of work are close enough that a citation is not a sufficient answer. Table 1 scores all of them on the same capabilities; what follows states what each row of that table means.

Sanchez-Lengeling et al. [9] is the nearest neighbour and the earliest. It built the first systematic attribution benchmark for molecular GNNs on tasks whose ground-truth substructure is known by construction, establishing that attribution methods should be scored against a known rationale rather than against how plausible a chemist finds the highlighting. MolSanity inherits that premise. What it adds is the faithfulness axis measured on the *same* molecules and the joint behaviour of the two. A benchmark that reports correctness alone cannot show that a faithfulness-based choice of attributor is the wrong choice, because it never makes that choice. The selection-consistency test of §3.3 makes it, and its outcome changes with the split.

MolFaith [13] is the closest work on the faithfulness axis and is larger than this study on that axis: eight attribution methods, two molecular representations and five architectures over roughly $14{,}000$ molecules, against the 31785 per-molecule records here. We make no claim to broader faithfulness coverage. Two of its findings bear directly on ours. First, it reports that some methods are inherently more faithful than others largely independently of architecture. We find the same stability from the other direction: faithfulness does not shift systematically across the 93 cells audited on both splits (median change 0.077). Two independent studies at different scales agreeing that faithfulness is a comparatively stable property of the method is the premise our result then acts on. The correctness ordering is also fairly stable across the split ($\rho = 0.821$ on MUTAG, 0.964 on MolMotif). Both rankings are stable; what neither study establishes is that they agree, and on our 5 molecular arms they do not — in either regime, and with the sign of their association varying by arm from -0.564 to $+0.786$. That is the distinction we would press: *stability of a proxy is not validity of a proxy*. A practitioner with no rationale labels who selects a method by faithfulness inherits a reproducible answer, not a correct one. Second, MolFaith reports that faithfulness varies markedly per molecule; our regime stratification locates part of that variation, since ground-truth localisation degrades precisely on confidently-wrong predictions while faithfulness on those same molecules *improves*.

GraphXAI [10] and *GraphFramEx* [11] are general-graph frameworks: GraphXAI supplies generators with exact node ground truth and an evaluation library, GraphFramEx systematises fidelity-style evaluation across explainers and tasks. Both are broader than MolSanity across graph domains and narrower within chemistry. MolSanity scores chemically meaningful units (SSSR rings, Murcko scaffold, BRICS fragments) rather than individual atoms, and stratifies by the split that governs molecular deployment. *DIG* [12] is a library rather than a protocol; we wrap its unfaithfulness metric, and its SubgraphX implementation is the coverage gap recorded in Limitations and threats to validity.

What is new here, and what is not. Ground-truth localisation, fidelity metrics, motif decompositions and scaffold splits all exist already. Two rows of Table 1 we do not find in any of the four: cross-checkpoint stability and calibration linkage. The contribution we would defend is neither of those rows alone, but the result they were built to support: that the two evaluation traditions above stand in no dependable relationship to one another — not in distribution, not under the scaffold shift molecular models are deployed into, and with an association that varies sharply between datasets under an identical audit protocol — and that no metric on the faithfulness side signals it.

3 The MolSanity audit

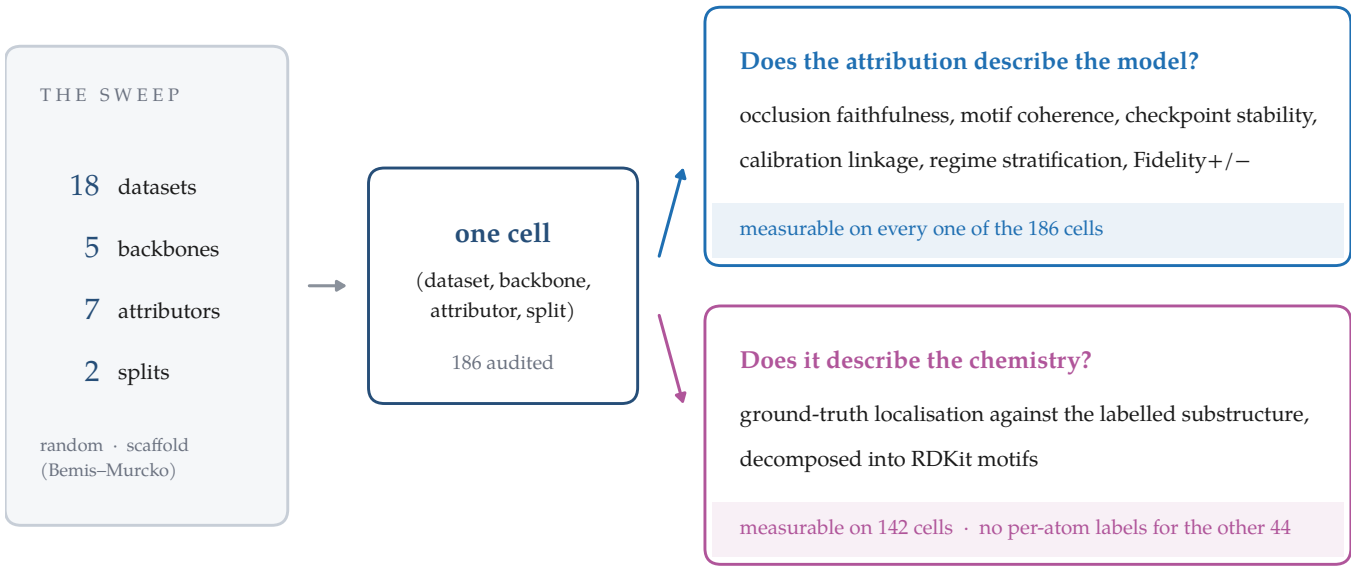
Figure 1 gives the shape of the audit before the definitions. The rest of this section makes each axis precise.

3.1 Notation

A molecule is a graph $G = (V, E)$ with node features $X \in \mathbb{R}^{|V| \times d}$ and edge features $Z \in \mathbb{R}^{|E| \times d'}$. A trained model f_θ maps G to logits $f_\theta(G) \in \mathbb{R}^C$ (classification) or a scalar (regression); t indexes the explained output. An *attributor* is a map $A : (f_\theta, G, t) \mapsto a \in \mathbb{R}^{|V|}$ assigning a real score to every atom.

Table 1: Where the audit sits relative to existing evaluation frameworks. ✓ = core capability, ∼ = partial or possible but not central, ✗ = not a focus. The contribution is the combination, not any single row: ground-truth validation and faithfulness both exist elsewhere, but not jointly with cross-checkpoint stability, calibration linkage and shift stratification over a motif-native decomposition.

Capability	GraphXAI	GraphFramEx	DIG	MolFaith	MolSanity
Ground-truth explanation benchmarks	✓ (synthetic)	∼	✓	✗	✓ (synthetic + motif-proxy)
Faithfulness / fidelity metrics	∼	✓	✓	✓	✓ (reproduced for comparability)
Molecular-motif-native audit (RDKit)	✗	✗	∼	∼	✓ (SSSR + Murcko + BRICS)
Occlusion–attribution agreement	∼	✓	∼	✓	✓
Cross-checkpoint stability	✗	✗	✗	✗	✓
Calibration linkage	✗	✗	✗	✗	✓
Scaffold-shift regime stratification	✗	∼	✗	✗	✓ (the core differentiator)
Multiple GNN backbones (agnostic)	∼	✓	✓	∼	✓ (GINE/GCN/GAT/MPNN/AttentiveFP)
Paired statistics (Wilcoxon, bootstrap CI)	∼	∼	∼	∼	✓
Across-seed variance reported per cell	✗	✗	✗	✗	✓ (3 seeds, 'SEED_VARIANCE.md')
Wraps canonical implementations (no re-impl)	✓	✓	✓	✓	✓ (Captum/PyG/RDKit)



The two answers disagree: ranking attributors by a model-side score picks the wrong one in 26 of 30 tests.

Figure 1: What the audit does, and the asymmetry it is built around. The sweep produces one *cell* — a (dataset, backbone, attributor, split) combination — as the unit of every claim in this paper. Each cell is scored on axes that ask whether the attribution describes the *model*, which are computable wherever an attribution is, and on one axis that asks whether it describes the *chemistry*, which needs per-atom labels and is therefore unavailable on every real molecular property dataset in the sweep. The model-side panel also carries Fidelity±, sparsity and the GraphFramEx characterisation score, computed on the same molecules for comparability. An evaluation framework that reports only the first column rates a cell in the lower-right of Figure 3 — faithful to the model, anti-aligned with the truth — as fine. Counts are read from the committed results, not annotated by hand.

RDKit decomposes G into a motif cover $\mathcal{M}(G) = \{M_1, \ldots, M_K\}$, $M_k \subseteq V$, from SSSR rings, the Bemis–Murcko scaffold [31] and BRICS-style fragments. Writing $a_v^+ = \max(a_v, 0)$, the motif-level score is

$$s_k = \sum_{v \in M_k} a_v^+, \qquad k = 1, \ldots, K. \tag{1}$$

Working at motif rather than atom granularity is what makes the audit chemically legible: a chemist reasons about a nitro group, not about atom 14.

For $S \subseteq V$ let $G \setminus S$ denote G with the features of S zeroed through the model's node-mask multiplier, $X' = X \odot (\mathbf{1} - \mathbb{1}_S)$, leaving the topology intact.

3.2 The five audit statistics

(i) Faithfulness. Occlusion of motif M_k changes the explained output by

$$\Delta_k \;=\; f_{\theta,t}(G) \;-\; f_{\theta,t}(G\backslash M_k). \tag{2}$$

An attribution is faithful when it ranks motifs as occlusion does, measured by the Spearman rank correlation over the K motifs of one molecule,

$$\rho_{\mathrm{occ}}(G) \;=\; \mathrm{Spearman}\big((s_k)_{k\le K},\,(\Delta_k)_{k\le K}\big) \in [-1,1]. \tag{3}$$

$\rho_{\mathrm{occ}} = 1$ means the attributor recovers the causal ordering exactly; $\rho_{\mathrm{occ}} = 0$ means it is uninformative about it; and $\rho_{\mathrm{occ}} < 0$ means it *inverts* it, which is strictly worse than an uninformative explanation. Undefined cases ($K < 2$, or zero variance in either argument) are recorded as missing, never as 0.

We also reproduce the field-standard fidelity pair [8]. With τ the 80th percentile of a^+ and $S_\tau = \{v : a_v^+ \ge \tau\}$ the salient set, and $p_t(\cdot)$ the softmax probability of the explained class,

$$\mathrm{Fid}^+ = p_t(G) - p_t(G\backslash S_\tau), \tag{4}$$

$$\mathrm{Fid}^- = p_t(G) - p_t(G\backslash \bar{S}_\tau), \tag{5}$$

higher Fid^+ and lower Fid^- being better, with sparsity $1 - |S_\tau|/|V|$. The GraphFramEx characterisation score [11] combines them as the weighted harmonic mean

$$\mathrm{char} \;=\; \frac{(w_+ + w_-)\,\mathrm{Fid}^+\,(1-\mathrm{Fid}^-)}{w_-\mathrm{Fid}^+ + w_+(1-\mathrm{Fid}^-)}, \qquad w_+ = w_- = \tfrac{1}{2}, \tag{6}$$

undefined outside $\mathrm{Fid}^\pm \in [0,1]$. That constraint matters for the regression cells, where Δ is taken in output space and $\mathrm{Fid}^\pm$ leaves that range.

(ii) Correctness. Where a node-level ground-truth mask $g \in \{0,1\}^{|V|}$ exists, correctness is the ability of a to separate the true substructure from the rest, scored by the area under the ROC curve of the min–max normalised attribution against g,

$$\begin{aligned}\text{GT-AUROC}(G) = \;&\Pr\left[\tilde{a}_u > \tilde{a}_v \,\middle|\, g_u{=}1,\, g_v{=}0\right] \\ &+ \tfrac{1}{2}\Pr\left[\tilde{a}_u = \tilde{a}_v\right],\end{aligned} \tag{7}$$

with u, v drawn uniformly from the positive and negative atoms. Chance is $\frac{1}{2}$; values below $\frac{1}{2}$ mean the attribution is *anti-aligned*, ranking the true motif systematically *below* the rest. When g is degenerate ($\sum_v g_v \in \{0, |V|\}$) the quantity is undefined and reported as such.

(iii) Coherence. Concentration and connectedness of the attribution mass: the Gini coefficient of a^+, the fraction of mass in the top 20% of atoms, the largest-connected-component fraction of S_τ in the induced subgraph, and the motif top-1 share $\max_k s_k / \sum_k s_k$. These are descriptive; Cross-task reliability shows they carry little reliability signal.

(iv) Stability. With θ^{early} an intermediate checkpoint of the same training run and s^{early}, s^{final} the corresponding motif scores (1),

$$\mathrm{stab}(G) \;=\; \mathrm{Spearman}\left(s^{\mathrm{early}},\, s^{\mathrm{final}}\right). \tag{8}$$

An explanation that reorders motifs between two checkpoints of one run is a weak basis for a chemistry decision.

(v) Regime and calibration. Logits are temperature-scaled on the validation fold [32]; let $c(G) = \max_j p_j(G)$ be the calibrated confidence and $y(G) \in \{0,1\}$ the correctness of the prediction. With $\tau_c = 0.8$, each molecule is assigned

$$R(G) = \begin{cases} \text{conf-correct} & c \ge \tau_c,\; y = 1, \\ \text{conf-error} & c \ge \tau_c,\; y = 0, \\ \text{borderline} & c < \tau_c. \end{cases} \tag{9}$$

Calibration linkage is the within-cell rank correlation between per-molecule calibration quality and reliability,

$$\rho_{\mathrm{cal}} \;=\; \mathrm{Spearman}\left(1 - |c - y|,\; \rho_{\mathrm{occ}}\right), \tag{10}$$

positive values supporting the hypothesis that better-calibrated predictions carry more trustworthy attributions.

3.3 The selection-consistency test

The audit's central question is operational rather than descriptive. A practitioner with no ground truth must choose an attributor using a faithfulness statistic; we ask whether that choice agrees with the one the ground truth would make.

Fix a *cell* $c = (\mathcal{D}, f_\theta, \pi)$, meaning a dataset, a trained backbone and a split, together with a set $\mathcal{A}$ of attributors, each audited on the same molecule sample $\mathcal{G}_c$. For attributor A write $\bar{m}(A) = |\mathcal{G}_c|^{-1} \sum_{G \in \mathcal{G}_c} m_A(G)$ for the cell mean of statistic m. Define the ground-truth-optimal and the faithfulness-optimal attributor

$$A^\star = \arg\max_{A \in \mathcal{A}} \overline{\mathrm{GT}}(A), \qquad \hat{A}_m = \arg\max_{A \in \mathcal{A}} \bar{m}(A), \tag{11}$$

for each faithfulness statistic $m \in \{\rho_{\mathrm{occ}}, \mathrm{Fid}^+, \mathrm{char}\}$. Two quantities follow. The *selection gap* is the paired per-molecule difference in correctness between the two choices,

$$\delta_m(G) = \mathrm{GT}_{A^\star}(G) - \mathrm{GT}_{\hat{A}_m}(G), \qquad G \in \mathcal{G}_c, \tag{12}$$

tested against $H_0 : \text{median}\,\delta_m = 0$ with a two-sided Wilcoxon signed-rank test on the molecules both attributors audited. The *selection-consistency coefficient* is the rank correlation between the two orderings over $\mathcal{A}$,

$$\rho_m^{\text{sel}} = \text{Spearman}_{A\in\mathcal{A}}\big(\bar{m}(A), \overline{\text{GT}}(A)\big) \in [-1, 1]. \quad (13)$$

$\rho^{\text{sel}} \approx 1$ means faithfulness is a sound surrogate for correctness in that cell; $\rho^{\text{sel}} \approx 0$ means it is uninformative; $\rho^{\text{sel}} < 0$ means it is actively misleading, in that ranking by faithfulness does worse than ranking at random. Because $\mathcal{D}$, f_θ and $\mathcal{A}$ are held fixed and only π varies between the two arms of each comparison, any change in ρ^{sel} is attributable to the split rather than to a change of dataset or model.

3.4 Estimation and inference

Cell means are reported with 95% percentile bootstrap intervals over molecules (2000 resamples, fixed seed). Attributor comparisons within a cell are paired on molecule identity and tested with the two-sided Wilcoxon signed-rank test; correlations are Spearman throughout, as all statistics here are ordinal in nature and several are heavy-tailed. Sample sizes are the audited molecules per cell (50–200), which is small enough that we treat individual p-values as descriptive and report them unadjusted for multiplicity, leaning on replication across cells rather than on any single test. Where a statistic is undefined for a molecule it is dropped from that statistic's mean only, and the retained n is reported alongside.

3.5 Implementation

The pipeline is staged and idempotent: each stage writes a marker keyed to a content hash of its configuration, every cell is isolated so one failure cannot abort a sweep, and trained checkpoints are content-addressed so an audit reproduces without retraining. Seeds, library versions, git revision and hardware are recorded per run. Attributors are the canonical implementations [17, 18], wrapped behind one interface rather than reimplemented.

4 Experimental setup

Datasets. Tier-1 provides the ground-truth anchor. Two arms are molecular with ground truth that is *exact by our own construction*: **MolMotif** and **MolMotifHard** take real drug-like molecules (BBBP and Tox21 respectively) and relabel them by the presence of a chemical substructure — an aromatic halogen and a carboxylic acid — so the substructure *defines* the label and its atoms are the node ground truth by definition rather than by assumption. MolMotifHard exists because MolMotif's halogen is a rare element the model can detect atom-wise, saturating the task; a carboxylic acid is built only from carbon and oxygen, so the model must learn the arrangement (§5.2). Two further molecular arms carry per-atom rationales published by others: **Benzene** and **FluorideCarbonyl**, from the attribution benchmark suite of Sanchez-Lengeling et al. [9] as repackaged by GraphXAI [10], whose labels we neither designed nor annotated and which therefore carry none of our own assumptions about what a rationale should be. **MUTAG** [29, 30] is molecular with the nitro proxy. Three arms have exact node labels but are not molecules: **SynthMotifs** and its larger twin **SynthMotifsXL**, generated offline (Barabási–Albert base plus house/cycle motifs); **BA-2Motifs** [5]; and **ShapeGGen** [10], a node-classification benchmark entered here at the graph level by taking each labelled node's k-hop enclosing subgraph, with k the message-passing depth. Tier-2 is MoleculeNet [27]: **BBBP** and **BACE** (classification), **ESOL**, **FreeSolv** and **Lipophilicity** (regression). Tier-3 covers chemistry-meets-medicine endpoints: **ClinTox**, **SIDER** and **Tox21** as single-task binary views of the MoleculeNet panels, and **DILI** (drug-induced liver injury) and **hERG** (cardiotoxicity) via the Therapeutics Data Commons [28]. TDC sets are featurised with the same PyG encoding as MoleculeNet so the backbones are drop-in. Credential-gated resources are out of scope and never worked around.

Backbones. GINE [20, 21], GCN [22], GAT [23], MPNN [24] and AttentiveFP [25], sharing a pooling head, calibration and splits.

Attributors. Integrated Gradients, Saliency, Guided Backpropagation and Input×Gradient through Captum [17]; GNNExplainer and PGExplainer through PyTorch Geometric [18]; SubgraphX [6] through DIG [12]. PGExplainer is classification-only by construction: its mask network trains against class logits, so it is absent from the regression cells. SubgraphX is scheduled on the ground-truth arms only — the cells where correctness can be measured — because its Monte-Carlo tree search is far more expensive per molecule than the gradient family (§9); on those arms it is a full column at the same sample size as every other attributor, not a reduced-budget approximation.

Splits. The primary split is a deterministic, class-aware Bemis–Murcko scaffold split [31], the protocol used to evaluate molecular property models under realistic generalisation [26] (whole scaffold groups move together, so there is no leakage, but every fold is guaranteed to contain every class), with a random split as the in-distribution reference. Imbalanced classification uses inverse-frequency class weighting and AUC-based model selection. Every

Table 2: The run ledger. Outcome of every cell the `full.yaml` sweep attempted, as recorded in `results/PROGRESS.md`. The *carried* column counts rows surviving from an earlier reduced-budget run because this run's attempt failed; it is zero, so every row in this paper is from the sweep reported here.

dataset	done	failed	skipped	carried
BA-2Motifs	36	0	0	0
BACE	12	0	0	0
BBBP	42	0	0	0
Benzene	42	0	0	0
ClinTox	12	0	0	0
DILI	6	0	0	0
ESOL	24	0	0	0
FluorideCarbonyl	42	0	0	0
FreeSolv	6	0	0	0
Lipophilicity	6	0	0	0
MUTAG	66	0	0	0
MolMotif	66	0	0	0
MolMotifHard	66	0	0	0
SIDER	12	0	0	0
ShapeGGen	42	0	0	0
SynthMotifs	66	0	0	0
Tox21	6	0	0	0
hERG	6	0	0	0
total	**558**	**0**	**0**	**0**

No cell failed and none was skipped.

comparison in this paper that is labelled "shift" versus "in-distribution" is these two splits on the same dataset.

Scale of the study. The sweep trains every (dataset, backbone) pair on both splits and audits every (dataset, backbone, attributor, split) combination, giving 558 audited combinations and 31785 per-molecule audit records in total. Audited samples are 50–200 molecules per combination, drawn from the held-out test fold. Exact software versions, hardware, seed and the per-combination outcome ledger are recorded with the released artifacts (Data and code availability).

Coverage. All 558 planned combinations completed: 558 done, none failed, none skipped (Table 2). Earlier sweeps of this study carried rows forward from a reduced-budget CPU run where a cell had failed; this one carries 0, so every number in this paper comes from the single run reported here. Collapsing the 3 seeds, the matrix holds 174 classification and 12 regression rows, of which 142 carry a node-level ground truth.

5 Results

Figure 2 states the paper's central result and, beside it, how far that result generalises. The rest of this section is the evidence behind those three panels: the selection test that turns the correlation into a concrete wrong recommendation (§5.2), what a practitioner can do instead when no ground truth exists (§5.5), the regimes in which attributions fail, and the arms and axes on which the audit found nothing, which we report at the same length as the ones where it did. Supplementary notes and the summary tables behind these sections are Additional file 1; the complete per-cell audit matrix every figure and quoted statistic is computed from is Additional file 2.

5.1 Faithfulness is not correctness

Table 6 lists every committed cell in which attribution *correctness* can be measured at all, and Figure 3 plots the joint distribution. Of the 142 cells with both metrics defined, 33 are anti-aligned with the ground truth (GT AUROC < 0.5) and 14 of those are simultaneously positively faithful to the model. That combination is the failure mode a faithfulness-only audit cannot see, because from the model's point of view nothing is wrong.

The clearest molecular instance is the MUTAG scaffold block of Table 6, where dataset, backbone and split are fixed and only the attributor varies. Guided Backpropagation, Saliency and Input×Gradient are the three *most* faithful attributors on that cell ($\rho = 0.616, 0.404, 0.499$) and the three *worst* at recovering the nitro motif (GT AUROC $0.013, 0.002, 0.079$). GNNExplainer, which the ground truth ranks best at 0.826, is less faithful (0.379) than all three. On this cell the two axes are not merely uncorrelated; they are inverted.

5.2 A fidelity-only ranking picks the wrong attributor, in both regimes

If faithfulness and correctness dissociate, then ranking attributors by a faithfulness metric, which is what an evaluation framework must do when no ground truth is available, can select the wrong method. 5 cell families let us test this cleanly: each is molecular, has a node-level ground truth, all 7 attributors audited on the same molecules, and both splits from this run, so the in-distribution and shift arms differ *only* in the split (Table 5, Figure 5).

MUTAG, under shift. The ground truth ranks GNNExpl. best (0.826). None of the 3 metrics finds it. Occlusion ρ ranks GuidedBP first, at 0.013 — sixth of the 7, and anti-aligned with the motif. Fidelity+ and the characterisation score both rank SubgraphX first, at 0.330, third. On paired per-molecule GT AUROC over the 53 shared molecules the median gap for the occlusion pick is 0.967 with $p < 0.001$, and $q < 0.001$ after controlling the false discovery rate over all 30 selection tests, so this contrast is

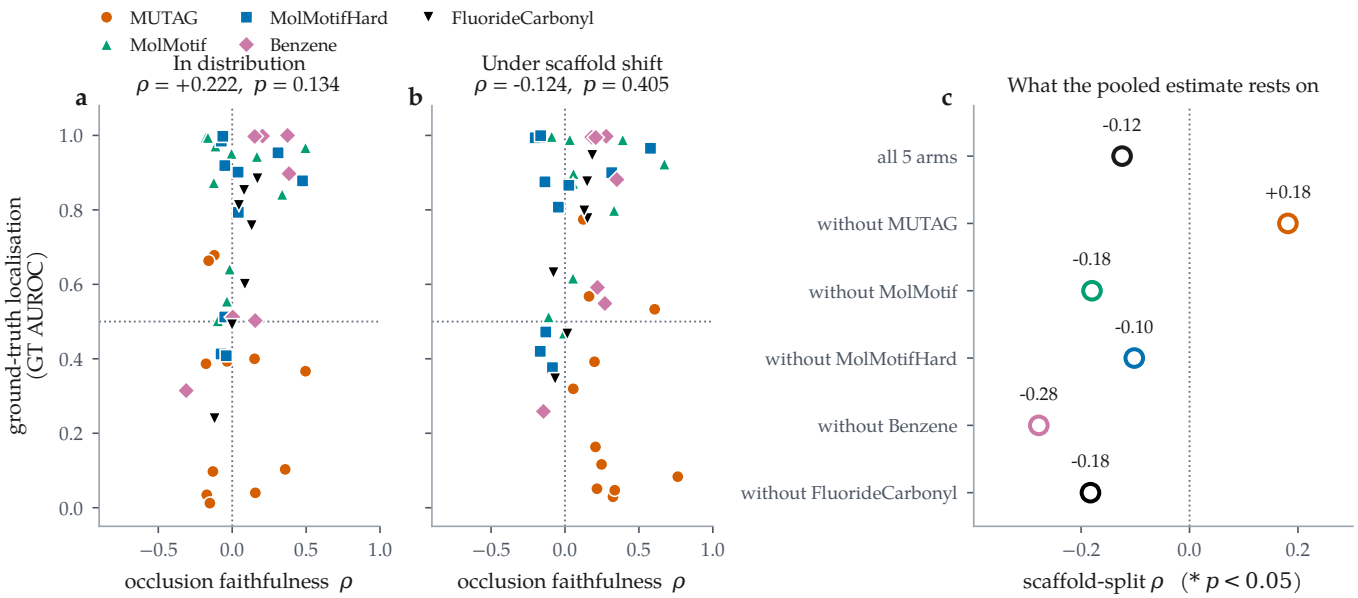


Figure 2: The result and its scope, together. Each point is one (dataset, backbone, attributor) cell of the 5 molecular ground-truth arms, averaged over 3 seeds; $n = 47$ cells. **(a)** In distribution and **(b)** under a Bemis–Murcko scaffold split, occlusion faithfulness carries no dependable information about ground-truth localisation: $+0.222$ ($p = 0.134$) and -0.124 ($p = 0.405$), neither distinguishable from zero. The horizontal line is chance localisation, the vertical line zero faithfulness. Note how much of each panel is populated at all — cells sit at every combination of high and low on the two axes, which is the dissociation itself. **(c)** What the pooled coefficient in (b) rests on. Filled markers would be significant at $p < 0.05$; all 5 leave-one-out fits are open, and dropping MUTAG reverses the sign to $+0.181$. This is not a robust effect being diluted by weak arms but a set of cell-specific effects that do not share a direction (§5.3). We put that panel here rather than in the limitations so the claim and its scope are read at the same time.

Table 3: Molecular regression cells. Occlusion faithfulness is computed in output space (predicted-value shift) rather than probability space. In 0 of the 12 committed regression cells the attribution–occlusion agreement is *negative*: the atoms an attributor ranks highest are not the atoms whose removal moves the prediction most. ESOL·GAT, ESOL·GCN, ESOL·GINE, FreeSolv·GINE, Lipophilicity·GINE run positive. Note the Fidelity± columns leave the $[-1, 1]$ range a probability-space fidelity would occupy (up to 7.27), which is why we report but do not interpret their sign. Bold marks the highest occlusion ρ per dataset (emphasis only). PGExplainer is classification-only and therefore absent by construction.

dataset	backbone	attributor	split	n	RMSE	MAE	R^2	motif top-1	occ. ρ	occ. top-1	Fid+	stab.
ESOL	GAT	IG	random	200	0.764	0.568	0.867	0.831	0.803	0.933	3.11	0.947
ESOL	GAT	IG	scaffold	200	1.039	0.799	0.762	0.863	**0.845**	0.973	7.27	0.955
ESOL	GCN	IG	random	200	0.947	0.737	0.797	0.854	0.423	0.650	−0.84	0.965
ESOL	GCN	IG	scaffold	200	1.071	0.867	0.747	0.887	0.437	0.767	−0.77	0.867
ESOL	GINE	GNNExpl.	random	200	0.833	0.629	0.842	0.849	0.634	0.832	−1.19	0.958
ESOL	GINE	GNNExpl.	scaffold	200	0.988	0.743	0.785	0.879	0.308	0.742	−0.40	0.890
ESOL	GINE	IG	random	200	0.833	0.629	0.842	0.865	0.683	0.830	−1.46	0.932
ESOL	GINE	IG	scaffold	200	0.988	0.743	0.785	0.897	0.409	0.753	−0.65	0.909
FreeSolv	GINE	IG	random	193	1.483	1.065	0.830	0.865	0.393	0.772	−0.69	0.862
FreeSolv	GINE	IG	scaffold	193	1.336	0.999	0.811	0.865	**0.421**	0.898	−0.61	0.703
Lipophilicity	GINE	IG	random	200	0.773	0.591	0.569	0.786	**0.556**	0.737	0.32	0.749
Lipophilicity	GINE	IG	scaffold	200	0.920	0.705	0.424	0.826	0.494	0.712	0.51	0.762

not an artefact of testing many of them. The faithfulness–truth rank correlations on this cell are -0.643 for occlusion ρ, -0.321 for Fidelity+ and 0.036 for characterisation. A framework ranking on faithfulness here does not find the best attributor, and the metric that ranks the field most nearly in reverse puts an anti-aligned attributor at the top of it.

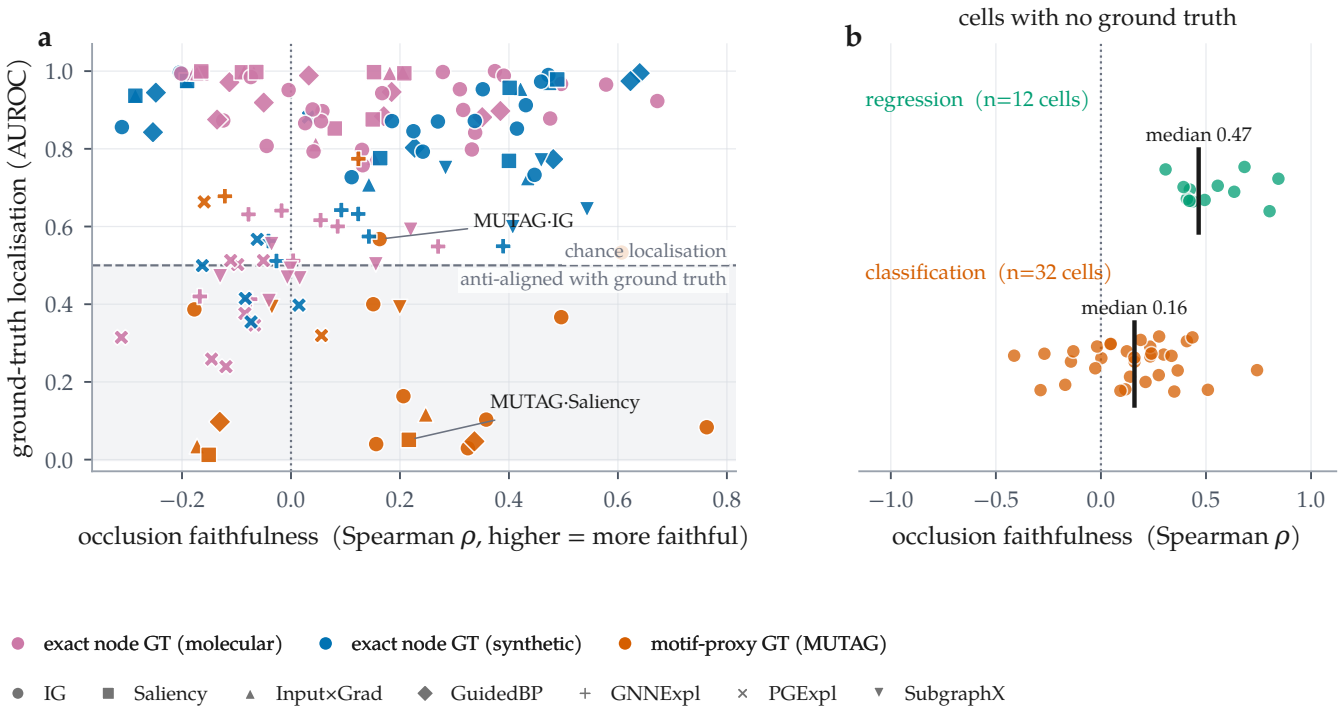


Figure 3: The two axes come apart, and for most cells only one of them exists. **(a)** Every committed cell carrying a node-level ground truth ($n = 142$), plotted as occlusion faithfulness against ground-truth localisation; colour separates the three kinds of ground truth — exact on real molecules, exact on synthetic graphs, and the MUTAG motif proxy — and marker distinguishes attributor. The shaded band is GT AUROC < 0.5: attributions there are anti-aligned with the true motif. 14 cells sit in the lower-right region, positively faithful to the model and anti-aligned with the truth. **(b)** The remaining 44 combinations, on real molecular datasets, have no published per-atom labels: only the horizontal axis is measurable for them.

MUTAG, in distribution. The random split does not rescue the ranking. All 3 metrics miss again: occlusion ρ selects GuidedBP at 0.037 against GNNExpl. at 0.858, a median paired gap of 0.978 over 58 shared molecules ($p < 0.001$, $q < 0.001$), and Fidelity+ and the characterisation score select IG and SubgraphX at 0.496 and 0.348. What separates the two regimes on this arm is not whether the pick is wrong but the rank correlation, which runs $+0.143$ in distribution against -0.643 under shift: in distribution the metric orders the attributors roughly correctly and still tops the list with the wrong one.

Across all 5 arms. 26 of the 30 selection tests are mismatches — 14 of 15 in distribution and 12 of 15 under scaffold shift. A fidelity-only ranking is wrong far more often than it is right, and it is wrong in *both* regimes; in-distribution evaluation is not the safe case. The regimes do differ, but in the cost of the mistake and in the opposite direction to the one we expected. Measured as the ground-truth localisation given up by choosing the metric's pick over the ground-truth-best attributor, the mean gap is 0.417 in distribution against 0.191 under shift (paired Wilcoxon over the 15 arm–metric pairs, $p = 0.004$). The narrowing sits entirely on the pick side: the ground-truth-best attributor localises at 0.946 and 0.952 in the two regimes, while the attributor the metric selects moves from 0.529 to 0.761. Under shift the metrics keep choosing the wrong attributor and happen to choose a less damaging one. Across mismatches the cost spans 0.002 to 0.821 AUROC, with a median of 0.369. A mismatch that gives up 0.002 and one that gives up 0.821 are the same entry in the "mismatch" column of Table 5 and are not the same result, which is why we report the gap beside the verdict throughout.

The pooled result: no reliable relationship in either regime. Any single arm gives a rank correlation over seven attributors, which is a weak instrument, so we pool. Over all 47 cells of the 5 molecular ground-truth arms, the faithfulness–correctness correlation is $+0.222$ in distribution ($p = 0.134$) and -0.124 under scaffold shift ($p = 0.405$). Neither is distinguishable from zero. Faithfulness itself rises across the split (0.049 to 0.132, $p = 0.005$) while ground-truth localisation does not move ($p = 0.539$), so the two axes move independently rather than together. The conclusion we draw from the pool is the null one, and it is the one that matters operationally: at this sample size a faithfulness score carries no dependable information about attribution correctness on molecular data, whichever regime it is measured in.

The per-arm coefficients show why no pooled estimate

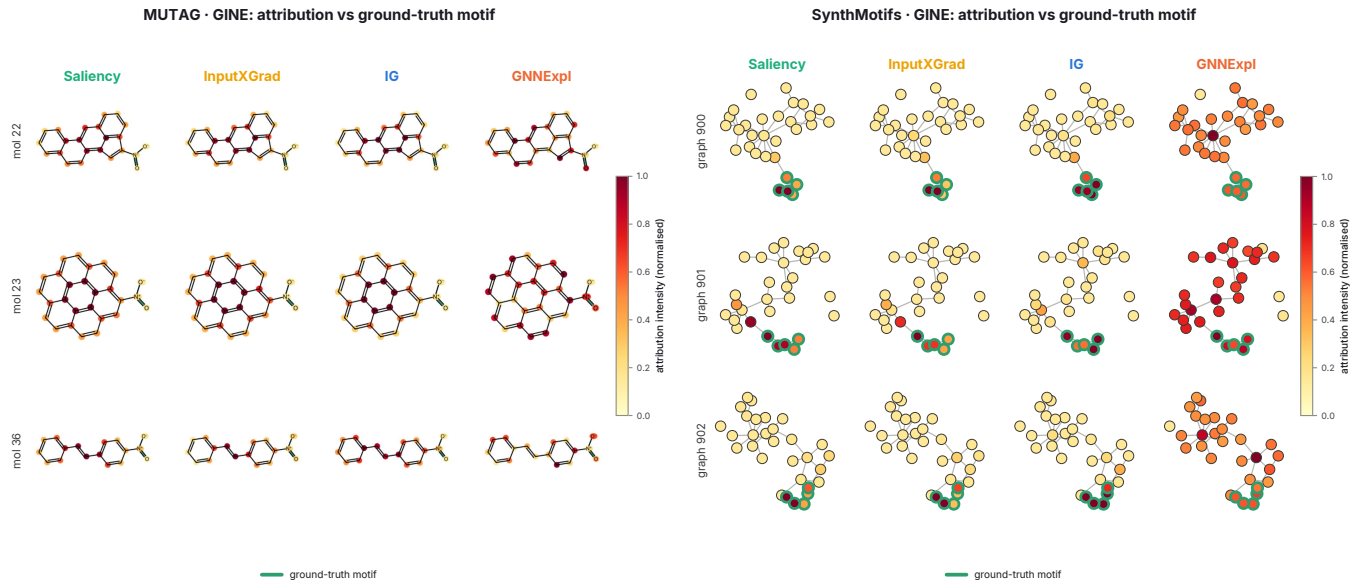


Figure 4: What the disagreement looks like on the structures themselves. Attribution intensity overlaid on the audited graphs for four attributors on the same molecules, rendered by the audit run: RDKit skeletal structures for MUTAG (left) and node-link diagrams for SynthMotifs (right), with the ground-truth motif outlined in green. The quantitative gap of Table 5 is visible here as a qualitative one: on the synthetic graphs the gradient attributors concentrate on the outlined motif while GNNExplainer spreads mass across the Barabási–Albert base, and on MUTAG the gradient methods saturate the fused aromatic system rather than the nitro groups that carry the mutagenicity label. SynthMotifs appears here as a structural illustration only: it is not a molecular dataset, so it has no Bemis–Murcko scaffold and takes no part in the shift contrast (§7). These panels are the pipeline's own output, reproduced unmodified.

should be trusted here. Under shift they run from -0.564 on MUTAG to $+0.786$ on FluorideCarbonyl — opposite signs, near the extremes of the coefficient's range, under an identical protocol differing only in the dataset. 3 of 5 arms fall between the two regimes and 2 of 5 are negative under shift, but the spread swamps the central tendency, and averaging over it produces a number that describes none of the five arms.

Leave-one-arm-out makes the same point quantitatively, and it is the first check a sceptical reader runs, so we run it. Removing each arm in turn from the shift pool: without MUTAG the correlation is $+0.181$ ($p = 0.290$, $n = 36$) — the sign reverses; without MolMotif, -0.180 ($p = 0.294$, $n = 36$); without MolMotifHard, -0.102 ($p = 0.554$, $n = 36$); without Benzene, -0.278 ($p = 0.082$, $n = 40$); without FluorideCarbonyl, -0.183 ($p = 0.259$, $n = 40$). 0 of the 5 fits is significant, and dropping a single arm is enough to flip the sign of the estimate. This is not a robust effect being diluted; it is a set of cell-specific effects that do not share a direction.

Restricted to the 3 arms that predate the two externally authored benchmarks, the same computation gives -0.356 at $p = 0.042$ over 33 cells — a significant collapse. Adding those two arms removes it (§5.8). The 5-arm result is the reported one. The structural reason the per-arm coefficients are so unstable is visible in the rankings themselves: on MUTAG 0 attributors exceed GT AUROC 0.9 and the 7 are spread over $0.002 - -0.826$, whereas on MolMotif under shift 4 of 7 sit above 0.9 within a few hundredths of each other. A Spearman coefficient over a group that is effectively tied is decided by whichever poor attributors happen to order correctly beneath it, and we would not draw a conclusion from any of the 5 arms on its own.

MolMotifHard exists because of this. MolMotif labels on halogen-aromatic presence, which a message-passing model can satisfy by detecting a rare *element* rather than an arrangement; MolMotifHard labels on a carboxylic acid, built from carbon and oxygen alone, so no single-element shortcut is available. It widens the usable range under shift (GT AUROC spanning $0.303 - -1.000$ against $0.476 - -0.998$) and leaves 3 rather than 4 attributors bunched above 0.9, without removing the bunching altogether, which is why it widens the evidence without settling the question by itself.

The arms that could not participate at all are the non-molecular ones. SynthMotifs, BA-2Motifs and ShapeGGen have exact node labels but are not molecules, so a Bemis–Murcko scaffold is undefined for them and their "scaffold" split comes back as an arbitrary deterministic partition (§7). The analysis therefore rests on the 5 molecular arms and their 47 cells: one chemically motivated proxy (MUTAG), two whose ground truth is exact because we built the label from the substructure (MolMotif, MolMotifHard), and two whose per-atom

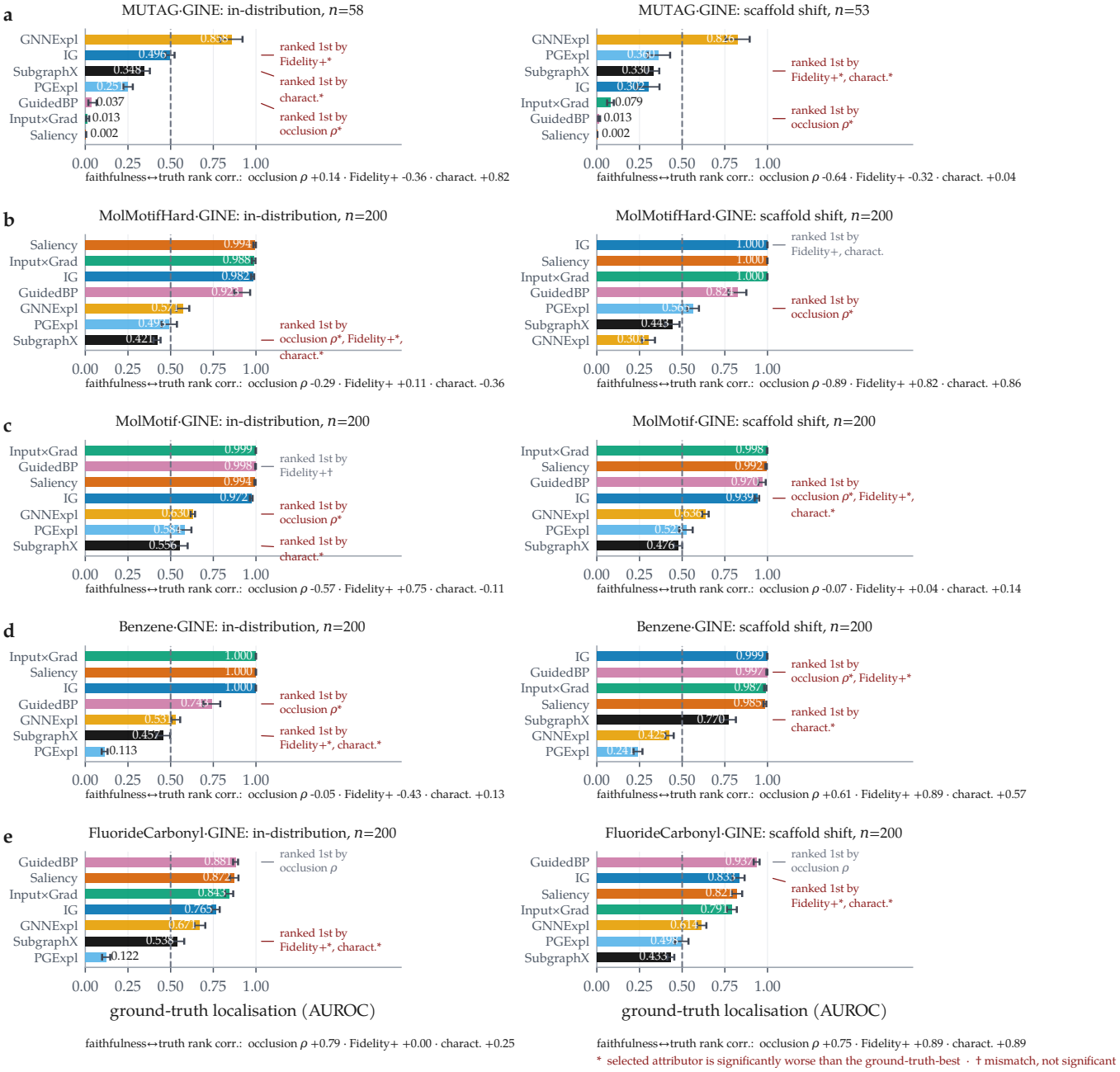


Figure 5: A faithfulness-only ranking selects the wrong attributor in 26 of 30 tests, in distribution as well as under shift. Ground-truth localisation per attributor, with bootstrap 95% confidence intervals over molecules. Within each row the dataset, backbone and attributor set are fixed and only the split changes; significant picks are marked *, picks statistically indistinguishable from the best are marked †. **Rows, top to bottom:** MUTAG (motif-proxy ground truth), where all 3 metrics mismatch in both regimes and under shift the occlusion measure selects GuidedBP, whose localisation (0.013) is anti-aligned with the motif; MolMotifHard and MolMotif (real molecules, exact node ground truth by construction), where the label *is* the substructure; and **Benzene** and **FluorideCarbonyl**, the Sanchez-Lengeling et al. [9] rationale benchmarks, whose per-atom labels we did not author. The mismatch is near-universal; what varies across rows is its cost, from 0.002 to 0.821 AUROC. Rank correlations between each faithfulness metric and ground truth are given below each panel.

rationales were published by Sanchez-Lengeling et al. [9] and which we neither designed nor labelled (Benzene, FluorideCarbonyl).

Two caveats. First, our own faithfulness metric fails alongside the field-standard ones. Occlusion ρ matches the ground-truth-best attributor on 14 in-distribution and 12 shift arms just as Fidelity+ and the characterisation score do, so we make no claim that MolSanity's occlusion measure is a better selector. *No* faithfulness metric in this experiment carries reliable information about which attributor is correct, which

argues for measuring correctness rather than for building a better proxy. Second, the 4 non-mismatches in Table 5 are not evidence that a metric worked: two are Fidelity+ and the characterisation score agreeing with each other on MolMotifHard under shift, and two are the same occlusion pick on both FluorideCarbonyl arms. Uniform-random selection among the 7 audited attributors would be expected to agree 4.3 times in 30 tests, so 4 is not above chance, and we do not read these rows as a positive result for any metric.

5.3 Reliability is a property of the cell

The preceding subsection reaches a null: over 47 molecular ground-truth cells there is no pooled faithfulness–correctness relationship in either regime. Under an identical protocol — same backbone, same 7 attributors, same audit code, same seeds — the per-arm coefficient under shift runs from -0.564 on MUTAG to $+0.786$ on FluorideCarbonyl, and the cost of a fidelity-only selection runs from 0.002 to 0.821 AUROC. Only the dataset changes.

A range is not by itself evidence that the arms differ. With 7–11 cells per arm, coefficients this far apart can arise from one underlying value, so we test it: Cochran's Q on the Fisher-transformed per-arm coefficients over the 5 arms gives $Q = 8.53$ on 4 degrees of freedom, $p = 0.074$, with $I^2 = 53\%$. That is moderate heterogeneity, and it does not reach significance. We therefore cannot claim that these arms have genuinely different coefficients, and we do not.

What the data do establish is narrower and sufficient for the practical point: no pooled coefficient here can be carried to a new cell. 0 of the 5 leave-one-arm-out refits survive at $p < 0.05$, and dropping a single arm reverses the sign (Figure 2c). Whether the arms truly differ or the estimate is merely too imprecise to pin down, a reader who quotes one number for "the" faithfulness–correctness relationship is quoting something this matrix does not support. This subsection shows the same instability along the other two axes of the cell, the backbone and the attributor, which is why we take the cell — (dataset, backbone, attributor, split) — to be the unit a reliability claim has to be indexed to.

When is the relationship positive?. The natural follow-up is whether the arms that come out positive share something the negative ones lack. Correlating the per-arm shift coefficient across the 5 arms against 3 properties of the arm gives: spread of ground-truth localisation within the arm, -0.500 ($p = 0.391$); its median level, 0.300 ($p = 0.624$); median faithfulness, 0.200 ($p = 0.747$). None approaches significance — the smallest p is 0.391 — and at $n = 5$ arms none could without an effect close to perfect. We therefore have no supported account of the sign, and prefer to say so than to offer one.

One pattern is visible, and we flag it as a conjecture rather than a result because it is confounded with everything else and cannot be tested at this n: the arms order by who authored the labels. The proxy arm is the most negative (MUTAG, -0.564), the two whose labels we constructed ourselves sit near zero, and the two whose per-atom rationales were published by others are both positive, the larger being FluorideCarbonyl at $+0.786$. If that ordering is real it says the association between faithfulness and correctness depends partly on how the ground truth was authored, which would make the quantity a property of the benchmark rather than of the attributor — a more uncomfortable finding than the heterogeneity itself. Separating it from coincidence needs molecular rationale benchmarks that do not currently exist (§7).

Figure 6 lays the two axes over the audited grid. Two further comparisons ask whether either axis is a stable property of the architecture or of the attributor, and the answer is more interesting than we expected: both are reasonably stable, which is what makes the instability of the *relationship* between them hard to see from any of the usual diagnostics.

Backbones. Holding the attributor fixed at Integrated Gradients and varying the backbone (Figure 7) gives a spread of 0.191 AUROC in distribution (GINE 0.985 down to GAT 0.793) and 0.186 under shift (GINE 0.994 down to GAT 0.807), comparable to the spread across attributors on the same dataset. Architecture is therefore a first-order factor in how well an attribution localises, and a report that omits it is under-specified. The ordering, however, is largely preserved across the split ($\rho = 0.900$ on MolMotifHard; 0.700–0.900, median 0.800, across the 3 molecular arms carrying all 5 backbones, all 3 positive): the architecture that localises the motif best in distribution mostly is the one that does so under shift. The attributor ordering is comparably robust ($\rho = 0.857$ over 7 attributors).

The choice of arm decides this result, which is worth stating because the arms are not interchangeable. Computed on SynthMotifs — exactly labelled, but not molecular, so its "scaffold" partition is the arbitrary deterministic one of §7 — the same correlation is $\rho = 0.100$, and would support the opposite conclusion. A rank correlation taken across a partition that is not a chemical regime is not evidence about shift, whatever it evaluates to, and we restrict this comparison to molecular arms for that reason.

The shift itself does not lower faithfulness. It would be tidy if scaffold shift simply made attributions less faithful. It does not. Across the 93 cells this run audited on both

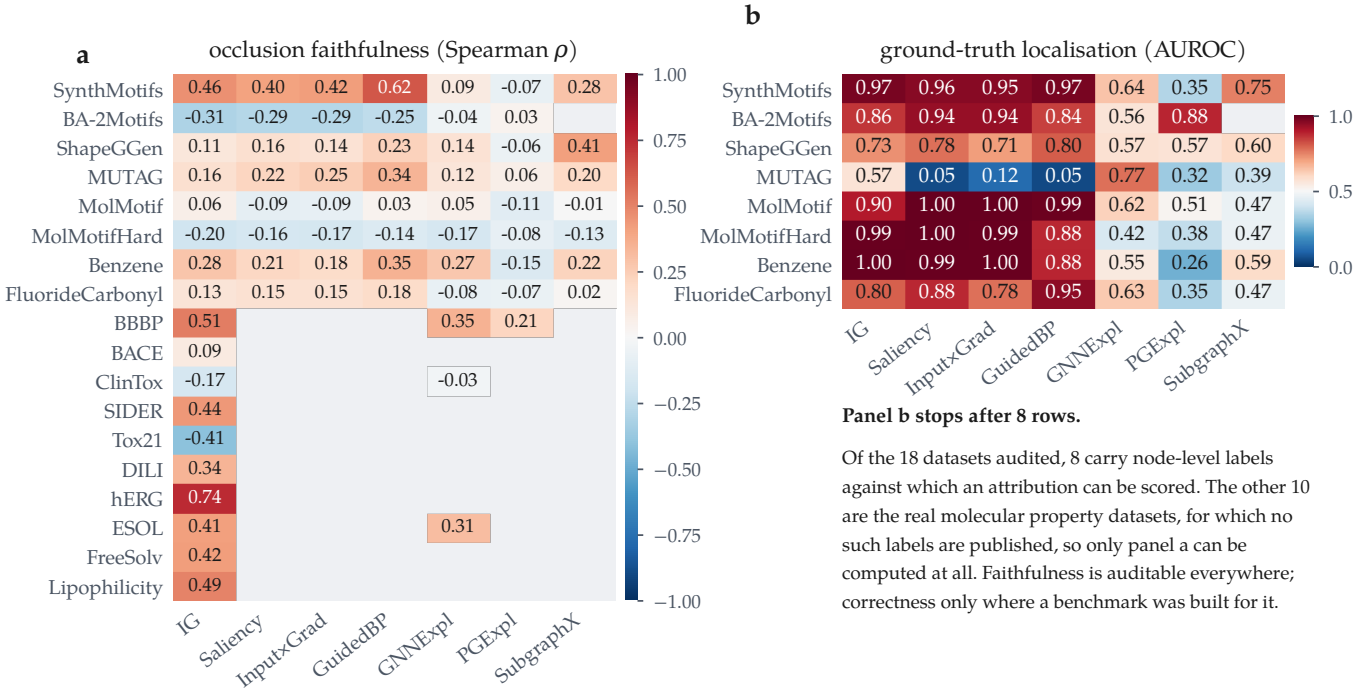


Figure 6: Reliability is a property of the (attributor, dataset) pair, not of the attributor. All committed GINE cells at the scaffold split. (**a**) Occlusion faithfulness over every audited dataset; blue = anti-faithful. (**b**) Ground-truth localisation, which exists only for the arms built to have it. Panel b is short because correctness is unmeasurable elsewhere, not because it scores badly there: no real molecular property dataset in the sweep publishes per-atom labels. Blank cells were not audited.

splits, occlusion faithfulness falls under the scaffold split in 36 of them, roughly half, with a median change of 0.077 (Figure 9b) — a coin flip, on the complete grid. Ground-truth localisation is equally unmoved across the split (0.660 to 0.658, $p = 0.539$). Both levels survive the shift; what neither level reveals is that they were never coupled to begin with (§5.2), which is why a faithfulness metric gives no warning that anything is wrong in either regime.

5.4 Does the model actually read the rationale?

The standing objection to every result above is Faber et al.'s [14]: scoring an attribution against a labelled substructure is misleading if the trained model does not use that substructure, in which case a low GT AUROC is a fact about the model rather than about the explanation. The objection is correct, and it is testable rather than merely arguable. For each molecule we occlude the ground-truth substructure and record whether the prediction collapses; if it does, the model demonstrably reads that substructure, and its attribution can be judged against it on the model's own behaviour.

The objection applies to more molecules than it does not. 8,085 of the audited molecules carry a rationale the model does not rely on, against 7,524 that it does, and the split is visible in the scores: mean GT AUROC is 0.781 where the model reads the rationale and 0.719 where it does not. A framework that ignored this distinction would be scoring the model under the guise of scoring the explanation, on roughly three molecules in five.

The number that settles it is what remains. Among the 7,524 molecules whose prediction demonstrably depends on the labelled substructure, 1,395 — a fraction of 0.185 — receive an attribution *anti-aligned* with it: the substructure the model provably uses is ranked among the atoms the attribution deems least relevant. On those molecules no appeal to an alternative rationale is available. The attribution misdescribes a model by that model's own behaviour, and faithfulness metrics computed on the same molecules do not flag it. We therefore treat the Faber objection as a partition to report rather than a defeater: it removes a majority of molecules from the correctness argument, and what survives it is still a large, unambiguous population of wrong explanations.

5.5 What to do when there is no ground truth

Every result above depends on rationale labels, which is exactly what a working medicinal chemist does not have. If the audit only says "your attributions may be wrong and you cannot check", it has diagnosed a problem and left the reader where it found them. The weaker but usable question is not *which attributor should I trust* but *on which molecules should I decline to trust this one* — a selective-prediction framing, transplanted from pre-

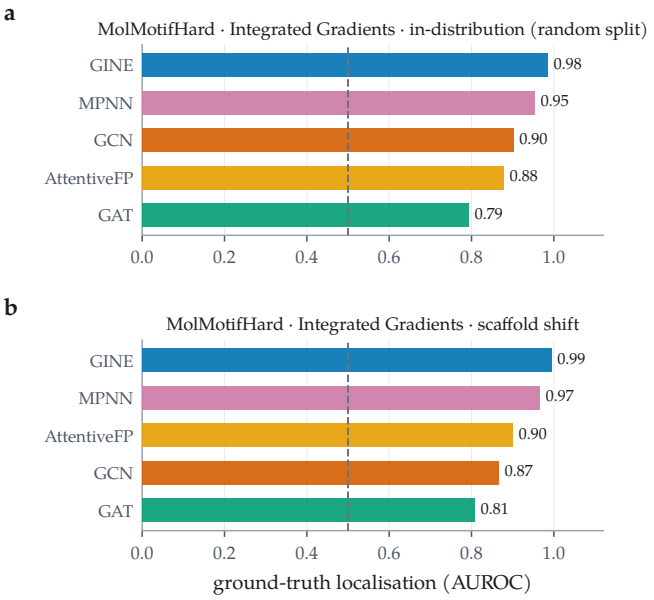


Figure 7: Which backbone localises best is a large effect, and a stable one across the split. Integrated Gradients held fixed; ground-truth localisation across the 5 backbones on MolMotifHard — a molecular arm with exact node labels, so its scaffold split is a chemical regime — in distribution (**a**) and under scaffold shift (**b**). Dashed line is chance. The two orderings correlate at $\rho = 0.900$ over 5 backbones, and at 0.700–0.900 across the 3 molecular arms carrying the full sweep. The gap between best and worst backbone is 0.191 AUROC in distribution and 0.186 under shift: the architecture matters, and it matters the same way in both regimes.

diction to explanation. We rank 5 candidate signals, all computable at inference time without any labels, by the gain in mean ground-truth localisation between full coverage and half coverage (Figure 8).

The best is occlusion faithfulness ($+0.056$ over $15,201$ molecules with ground truth). The rule worth stating would be an operating point at which at most 10% of retained molecules carry a below-chance attribution. *No coverage on this curve reaches it.* The below-chance share starts at 0.216 at full coverage and bottoms out at 0.145 at 37% coverage ($5,624$ molecules retained, mean localisation 0.814 against 0.748 at full coverage), then worsens again as coverage falls further. Abstention on the best available label-free signal therefore removes roughly a third of the below-chance tail at the cost of discarding most of the molecules, and cannot remove it. We report the rule with that ceiling attached rather than quoting the lift alone: a practitioner adopting it should expect to reduce the failure rate, not to reach a target. We state the rule against the below-chance share rather than the mean deliberately: pooled mean localisation is already tolerable at full coverage, so a rule stated against it would look successful while leaving the failures untouched. The tail is what a deployment gets hurt by, and the tail is what

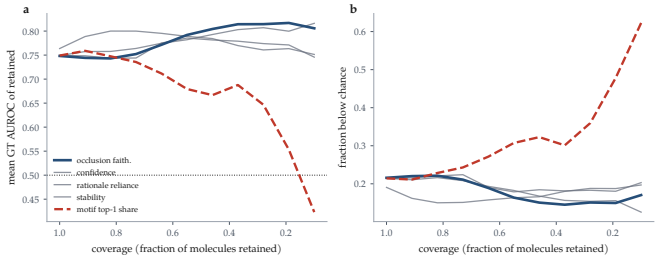


Figure 8: Coverage–reliability curves for abstention signals, over the $15,201$ audited molecules that carry ground truth. Every signal is computable at inference time without rationale labels. Moving left to right is abstaining on more molecules. (**a**) Mean ground-truth localisation of the retained set; the dotted line is chance. (**b**) The fraction of the retained set whose attribution is *below* chance — the quantity a deployment is hurt by, and the one the pooled mean in (a) hides. The best signal (occlusion faithfulness) is highlighted; motif top-1 share is dashed because its lift is negative, so abstaining by it makes the retained set worse.

this moves — partially.

Why this is not the argument of §5.2 in reverse. The signal that ranks best here is occlusion faithfulness, and §5.2 says faithfulness is not a dependable guide to correctness. These are different claims about different objects, and the distinction is load-bearing. §5.2 asks whether faithfulness ranks *attributors* against ground truth — a between-method question, answered per cell, and answered no. This asks whether, for one fixed attributor, faithfulness separates the *molecules* on which that attributor happens to localise well from those on which it does not — a within-method, per-molecule risk question. A signal can carry molecule-level risk information while being useless for choosing between methods, and on this evidence faithfulness does exactly that. It is also fitted where labels exist and evaluated there, so it is a benchmark-internal result: we do not claim the threshold transfers to a dataset with no ground truth, which is the setting that motivated it.

The ranking's negative result is as useful as its positive one. 1 of the 5 signals has *negative* lift: motif top-1 share (-0.082). Abstaining on molecules whose attribution mass is spread across many motifs — keeping those that concentrate on a single chemically coherent one — makes the retained set *worse*. That is the intuition a chemist is most likely to apply by eye, and on this data it is anti-predictive. An audit that reported only the working signal would have left that trap in place.

What this rule does not license. These curves are fitted where ground truth exists. Applying them to a dataset where it does not assumes the signal–reliability relationship transfers, and the central finding of this paper is that one such relationship — faithfulness to correctness

— does not survive a change of split. We therefore offer the rule as a calibrated starting point that a group can re-fit on its own labelled subset, not as a constant of nature. Stating that constraint is not hedging: it is the same standard we hold the faithfulness metrics to.

5.6 Where attributions fail: regimes and calibration

With per-molecule records available, two audit axes that previous drafts could only define are now measurable (Figure 10).

Regime stratification. Pooling the 31785 audited molecules by regime gives the audit's most directly actionable result. Ground-truth localisation is highest on confident-correct molecules (0.769, $n = 10722$), drops to 0.706 on borderline ones ($n = 4612$), and is lowest, at 0.681, on *confident-error* molecules ($n = 275$). It stays above chance throughout, so the attribution does not invert wholesale; it degrades exactly where a deployment can least afford it, in the regime where an explanation is most likely to be believed.

The two cheap proxies both point the wrong way there. Occlusion faithfulness on confident-error molecules is 0.325, markedly *better* than the 0.142 of confident-correct ones, and cross-checkpoint stability is 0.802 against 0.750 — also no worse. A practitioner screening explanations by faithfulness or stability would rank the least correct group at or above the most correct one. The confident-error estimate rests on 275 molecules and is reported with that n throughout; it is suggestive rather than precise.

Calibration linkage. The framework's hypothesis is that better-calibrated predictions carry more trustworthy attributions. Computed per cell it is supported on balance (median $\rho = 0.144$ across 174 classification cells, 85 significantly positive against 29 significantly negative), but the effect is cell-specific rather than universal. Pooling every molecule into a single correlation instead yields 0.074, roughly a quarter of the per-cell median and indistinguishable from no relationship at all. The pooled number is an artefact of mixing cells with different confidence distributions, and we report it only to warn against it: an evaluation that reports one aggregate correlation would conclude the linkage does not exist, when per cell it does.

5.7 Cross-task reliability

Tables 3, 10 and 11 give every audited combination on a real molecular dataset. Three cross-task patterns hold, and one apparent pattern turns out to be an artefact.

Occlusion faithfulness needs a different operator on regression. The obvious extension of the classification operator to regression is mis-specified, and it fails in a way that mimics a finding. If the attribution is clipped at zero — keeping only atoms that push the prediction *up* — while the occlusion effect keeps its sign, the two sides are no longer measuring the same quantity. A motif driving a solubility prediction strongly down then scores as unimportant on the attribution side and dominant on the occlusion side, which alone is enough to produce systematic negative faithfulness on cells whose predictive performance is perfectly well behaved (R^2 from 0.747 to 0.867 on ESOL). We flag it because a uniformly negative column invites a substantive reading, and the substantive reading would be wrong.

The operator used here ranks by magnitude on both sides, reports the GraphFramEx characterisation score as undefined rather than clipping a standard-deviation-space shift into $[0, 1]$, and adds a bounded statistic giving the share of the total occlusion effect carried by the salient atoms. Under it the pathology is absent: 0 of the 12 regression cells show negative occlusion faithfulness, with median 0.465.

Coherence is high nearly everywhere, and says little. Motif top-1 share lies between 0.060 and 1.000 across the 138 molecular cells, with 91 of them inside a narrow 0.6–0.9 band, largely independent of whether the same cell is faithful or anti-faithful. Concentration on a single motif is easy to achieve and, on this evidence, close to uninformative about reliability. It belongs in the audit as a descriptive statistic, not as a quality score.

Stability is high but not uniform. Median cross-checkpoint stability over 176 cells is 0.829, with a minimum of -0.101, a cell whose early and final attributions are essentially uncorrelated. By attributor, GNNExplainer is the most stable (median 0.938) and Integrated Gradients sits at 0.873 over 84 cells; the per-attributor counts are very unequal in this run, so these medians are descriptive rather than a controlled comparison.

Table 4 gives the paired attributor contrasts behind Section A fidelity-only ranking picks the wrong attributor, in both regimes, computed from the per-molecule records.

5.8 Negative and degenerate results

We record what did not work alongside what did.

A benchmark whose node labels are not where a loader looks for them. BA-2Motifs exists to carry exact node labels, but PyG's `BA2MotifDataset` exposes only node features, an edge index and a graph-level label: any extractor reading a per-node field finds nothing, and the arm

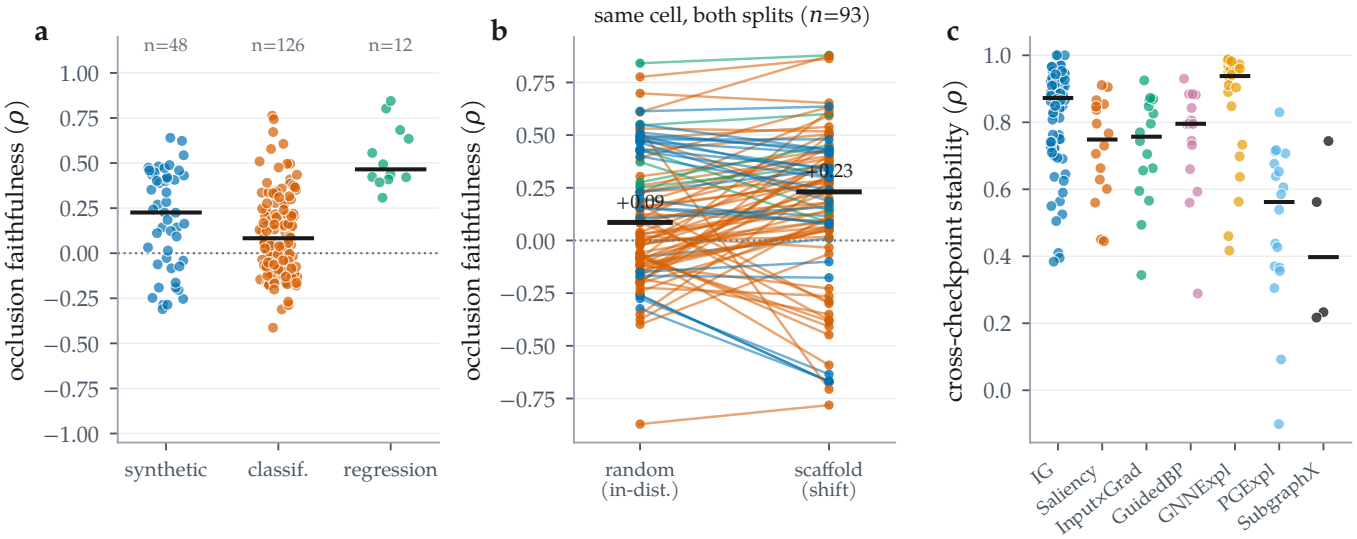


Figure 9: Distributions across all audited combinations. (**a**) Occlusion faithfulness by dataset regime; bars are medians. (**b**) The same cell on both splits, joined: faithfulness falls under the scaffold split in only 36 of 93 cells (median change 0.077), so shift does not systematically reduce faithfulness. (**c**) Cross-checkpoint stability by attributor; the mask-learning methods are not uniformly more stable than the gradient ones.

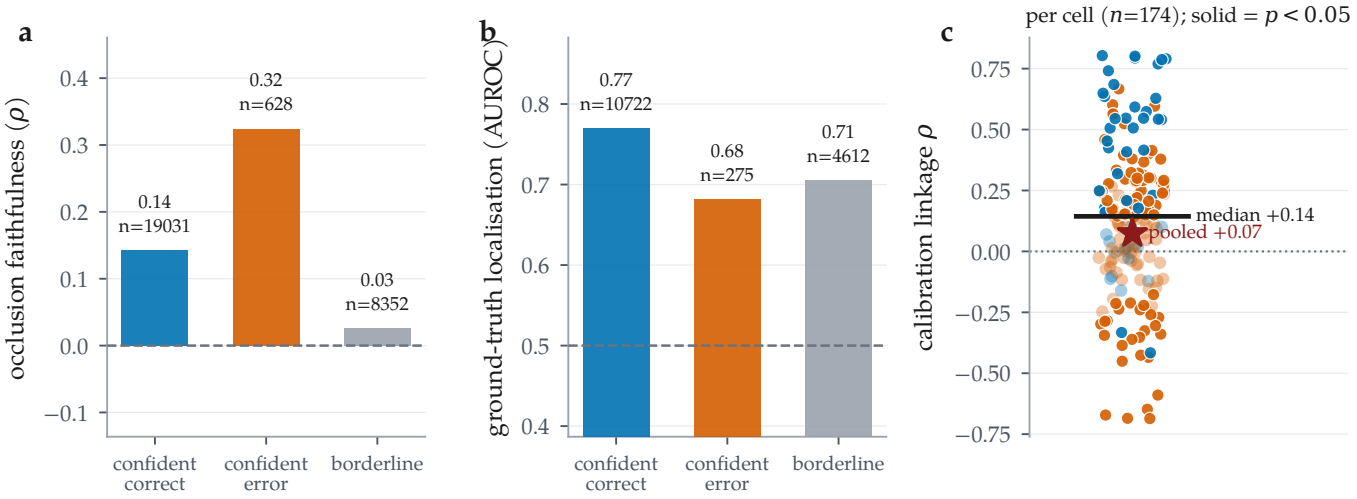


Figure 10: Regime stratification and calibration linkage, over the 31785 per-molecule audit records. (**a**) Occlusion faithfulness and (**b**) ground-truth localisation by confidence/correctness regime; n is printed per bar because ground truth exists for only a minority of molecules. Localisation on confident-*error* molecules degrades (0.769 to 0.681, $n = 275$) while occlusion faithfulness on the same molecules *improves* (0.142 to 0.325). (**c**) Calibration linkage computed per cell (solid = $p < 0.05$): the median cell is clearly positive, but pooling all molecules into one correlation attenuates it to near zero, a Simpson's paradox and a caution against reporting the pooled number.

silently lands in the no-ground-truth block despite being a ground-truth benchmark. The labels are recoverable from the released node ordering, in which the injected motif is appended after the Barabási–Albert base. We recover them and verify the recovery structurally, requiring the induced subgraph on those nodes to be a house or a five-cycle before the mask is used, cross-checked against a generator producing the same construction with independent labels. The mask is refused rather than guessed when the check fails, so a loader that changes the ordering degrades to no-ground-truth rather than inventing one.

BA-2Motifs accordingly carries node-level ground truth throughout and appears in Table 6 as a ground-truth arm. It is not molecular, so it takes no part in the shift contrast for the same reason SynthMotifs and ShapeGGen do not.

A perfect classifier with anti-faithful attributions. The highest-AUC molecular cell whose attributions are anti-faithful is Benzene · GINE on the scaffold split, at accuracy 0.925, ROC-AUC 1.000 and ECE 0.048: accurate, well calibrated, and explained by an attribution that ranks the motifs of these molecules against their causal order

($\rho = -0.146$). Predictive quality, calibration and explanation reliability are three different axes, and a report carrying the first two says nothing about the third.

A weak model can be faithfully explained. The converse also appears. SIDER · GINE on the scaffold split reaches $\rho = 0.436$ — among the most faithful cells in the matrix — on a model at ROC-AUC 0.627 and accuracy 0.555. Faithfulness certifies that the explanation describes the model; it says nothing about whether the model is worth describing. Any reliability report that omits the predictive metrics can be satisfied by this cell.

Accuracy is the wrong headline on imbalanced sets. The clearest case is Tox21, where accuracy runs 0.940–0.942 against AUC 0.745–0.820: a number that looks like a solved task and a number that does not, on the same models. Across the run, 37 committed rows report test accuracy ≥ 0.95. We report AUC alongside accuracy throughout for this reason, and the lowest-AUC molecular cell in the run (SIDER · GINE, AUC 0.627 at accuracy 0.555) is where accuracy alone would mislead most.

Two arms that remove the pooled shift effect. We added **Benzene** and **FluorideCarbonyl**, the Sanchez-Lengeling et al. [9] attribution tasks as repackaged by GraphXAI [10], precisely because they are molecular, exactly labelled, and *not constructed by us* — corroboration the three self-built or proxy arms could not provide. They removed the effect they were added to corroborate.

Restricted to the 3 arms that predate them, the pooled faithfulness–correctness correlation over 33 cells runs from $+0.009$ in distribution ($p = 0.962$) to -0.356 under scaffold shift ($p = 0.042$) — a significant collapse. Over all 5 arms and 47 cells the same computation gives -0.124 at $p = 0.405$. The effect does not shrink; it stops having a direction, because Benzene and FluorideCarbonyl both come out *positive* under shift. Both figures are computed from the same run by restricting the arm set, so they differ in which arms enter and in nothing else. A preprint of this work reported the 3-arm figure; the 5-arm figure supersedes it.

An exclusion criterion is only as good as its reason, and it must be discarded when the reason expires. These two arms are molecular but reach the audit as graph archives, and a loader that does not carry the SMILES the archives also contain leaves Bemis–Murcko with nothing to work on: the scaffold split then comes back degenerate, every molecule in its own bucket, and pooling it into a shift analysis would inject a deterministic partition into a regime contrast. The splitter reports `frac_grouped` for exactly this reason and flags the condition. Once the SMILES are carried and the partitions are real, the only justification for holding the arms out is gone, and holding them out anyway would be selecting arms on their result.

Two dependencies that are harder to install than to use. SubgraphX (via DIG) and ShapeGGen (via GraphXAI) both require the compiled `torch_sparse`/`torch_scatter` extensions, for which no wheel exists at the pinned torch version. Neither is genuinely unavailable: the extensions build from source, DIG then installs, and GraphXAI installs from a source checkout rather than from a release. Both are integrated here — SubgraphX as a full attributor column on every ground-truth arm, ShapeGGen as its own cells. We note the build path because a missing wheel is easily read as a hard block, and a search-family attributor or a benchmark dropped on those grounds is a coverage gap that no reader of the resulting table can see.

Two limitations survive that fix rather than being resolved by it. SubgraphX is two to three orders of magnitude more expensive per molecule than the gradient family, which is a selection pressure on which attributors get benchmarked at all and which we discuss in §7. ShapeGGen is a node-classification benchmark on a single large graph, whereas every axis of this audit is defined per molecule at the graph level; it enters the audit as a graph-level task by construction, and its scaffold split is degenerate for the same reason the other non-molecular arms' are.

6 Discussion

What the audit buys. The single most useful thing MolSanity produces is not a score but a *negative certificate*. For 44 of the 186 committed rows, covering every real molecular dataset in the sweep, attribution correctness cannot be measured at all, because no per-atom explanation labels exist. Those cells can be audited for faithfulness, coherence, stability and calibration, and this paper's central result is that none of those is a reliable stand-in for correctness in either regime. Making that gap explicit is more useful than filling it with a metric that looks like correctness and is not.

How attributions should be reported. Four practices follow directly from the matrix. (i) Report the cell, not the attributor: on the 5 families where we can check, the same protocol gives per-arm faithfulness–correctness correlations from -0.564 to $+0.786$ under shift, so "method X is reliable" is not a claim the evidence supports at the level of a method. (ii) Report the backbone: holding the attributor fixed, the choice of architecture moves ground-truth localisation by 0.125–0.538 AUROC, so "IG on a GNN" is under-specified — even though, unlike the faithfulness–

correctness relationship, the backbone ordering itself is stable across the split. (iii) Report the predictive metrics next to the explanation metrics, because the accurate, well-calibrated Benzene · GINE cell carries anti-faithful attributions ($\rho = -0.146$ at ROC-AUC 1.000). (iv) Do not pool: the calibration-linkage correlation attenuates from a per-cell median of 0.144 to 0.074 when cells are pooled.

Why reliability is regime-, backbone- and dataset-dependent. A faithfulness metric asks about the model; a ground-truth metric asks about the chemistry; they agree only when the model has learned the chemistry. That coincidence is engineered into synthetic benchmarks and is not guaranteed on molecular data, particularly out of distribution, which is the situation Faber et al. [14] identify as making ground-truth comparison hazardous in general. Attribution evaluation inherits the model's inductive limitations, so the correct object of study is the cell, (dataset, backbone, attributor, split), and not the attributor alone.

7 Limitations and threats to validity

Ground truth is scarce, and each arm of it is compromised differently. 5 molecular arms carry the argument, at $7--11$ paired cells each. None is unimpeachable, and they fail in different directions, which is part of why the per-arm coefficients disagree so sharply.

MUTAG uses a chemically motivated nitro/aromatic-nitro SMARTS proxy rather than annotator labels: it encodes an assumption about which substructure ought to matter, and the trained model may use a different, equally valid rationale [14]. We answer that objection empirically rather than by assertion (§5.4), but it remains a proxy, at $n = 53$ audited molecules per arm — the smallest of the 5. MolMotif and MolMotifHard are molecular *and* exactly labelled by construction, since the label is presence of a substructure, but their attributor rankings are *bimodal*: under shift 4 of 7 on MolMotif and 3 on MolMotifHard sit above GT AUROC 0.9 within a few hundredths of one another, so a 7-point rank correlation on either arm is carried by a handful of comparisons among effectively tied methods. We built MolMotifHard to widen that range and it does, but not enough to make either arm decisive alone. Benzene and FluorideCarbonyl are the only arms we neither designed nor labelled, which is why we weight them heavily, but they are binary substructure-detection tasks too, and a rationale that is a single functional group is an easier localisation problem than the multi-centre rationales that carry most real ADMET endpoints.

The arms that might have corroborated further cannot. SynthMotifs, BA-2Motifs and ShapeGGen have exact node labels but are not molecules, so a Bemis–Murcko scaffold is undefined for them: our splitter reports every graph in its own scaffold bucket, which makes their "scaffold" split an arbitrary deterministic partition rather than a chemical shift, and we exclude them from the contrast rather than let a degenerate partition stand in for one. No public dataset is simultaneously molecular, exactly labelled and *graded* in difficulty; constructing one is the clearest next step for this line of work. Until it exists, the honest reading of the pooled estimate is that it is uninformative, and the per-cell values are the reportable quantity.

A dataset that was not scored, and now is. For two sweeps BA-2Motifs contributed faithfulness only, appearing in the no-ground-truth block despite being a synthetic benchmark designed to carry exact labels, because our extractor did not read the ordering those labels are encoded in. The recovery is in this sweep and BA-2Motifs is scored on the correctness axis throughout. We leave the episode in the record because the failure was silent: a dataset whose entire purpose is exact ground truth sat in the *no*-ground-truth block for two runs without anything flagging the contradiction.

Three seeds, one split rule, and one exception. Each cell is run at 3 seeds, and every figure in the results matrix and in the pooled analysis is an across-seed mean. The selection tests (§5.2, Table 5, Figure 5) are the exception, and it is a consequence of the design rather than an oversight: the split is a function of the seed, so different seeds audit different molecules, and a *paired* per-molecule test requires the same molecules on both sides. Those tests are therefore computed within a single seed's records. The point estimates they report are correspondingly noisier than the matrix rows: on the MUTAG shift arm the ground-truth-best attributor scores 0.826 in the audited seed against an across-seed mean of 0.774 ± 0.086. The ordering is the same and the mismatch conclusion does not depend on the seed, but the number does, and we report both rather than only the one that supports the claim more comfortably. The full per-cell across-seed spread is Additional file 3. What is *not* varied is the split rule itself: each regime is one deterministic partition, so no interval here covers the variance of choosing a different scaffold-split implementation or ratio. Per-cell molecule counts are also modest where the dataset is small — $n = 53$ on the MUTAG scaffold fold, which is the entire held-out set rather than a sample from it — so individual contrasts remain underpowered even at 3 seeds, which is why the central claim is made on the pooled estimate rather than on any single cell.

Occlusion as a counterfactual. Zeroing node features takes the graph off the data manifold; a motif whose removal barely moves the output may be redundant rather than unimportant. It bears most visibly on the regression cells, whose output-space Fidelity± values are shifts in units of the training target's standard deviation and so leave the $[-1, 1]$ a probability-space fidelity occupies; they are comparable within a column and not across tasks.

Proxy ground truth. MUTAG's mask is a nitro/aromatic-nitro SMARTS proxy, not annotator labels. It encodes an assumption about which substructure ought to matter, and the trained model may use a different, equally valid rationale [14].

Coverage of attributors. The search family is represented by SubgraphX on the ground-truth arms only, because its cost per molecule is far above a gradient attributor's (§9). Conclusions that pool over the whole matrix therefore weight the gradient family more heavily than a cost-blind design would, and any claim about search-based attribution outside the ground-truth arms is unsupported here. The per-cell scores behind this comparison, including the field-standard Fidelity±, sparsity, the GraphFramEx characterisation score and the PyG/DIG unfaithfulness metric computed on the same molecules, are Additional file 4. This is a real coverage limit, and it is the same selection pressure — expensive attributors get benchmarked less — that the audit is meant to make visible rather than inherit silently.

Multiplicity. The selection tests carry Benjamini–Hochberg adjusted q values over the family of 30 tests, and the attributor contrasts carry them over the family within each cell block, so the headline mismatches are false-discovery-rate controlled rather than descriptive. What the adjustment cannot repair is power: at $n = 53$ molecules per arm a contrast that fails to reach significance is weak evidence of no difference, not evidence of none.

8 Conclusion

Molecular GNN attributions are audited today for whether they describe the model. Across the committed audit matrix we find that this is not the same question as whether they describe the chemistry, and that answering the first tells you very little about the second. On the 5 cell families where node-level ground truth, 7 attributors and both splits coexist, a faithfulness-only ranking selects an attributor other than the ground-truth-best one in 26 of 30 tests, in distribution as readily as under scaffold shift, at a cost ranging from 0.002 to 0.821 AUROC. MolSanity's own faithfulness metric is no exception, and we do not offer it as a better selector. The two remaining findings do different jobs and should not be merged. The pooled coefficients ($+0.222$ and -0.124 over 47 cells, neither significant) *fail to establish* a dependable global relationship: there is no faithfulness–correctness law here that a practitioner can appeal to. That is a weaker statement than proving none exists, and we intend the weaker one. The per-arm spread (-0.564 to $+0.786$ under an identical protocol) shows *why no pooled figure should be transported* even if a future study finds one: whether the arms genuinely differ — which 5 arms cannot settle, $Q = 8.53$, $p = 0.074$ — or the estimate is simply too imprecise, the quantity is not pinned down across the cells it would have to be applied to. Every leave-one-arm-out refit is non-significant and dropping one arm reverses the sign. A reliability claim therefore has to be indexed to the cell; a single headline number about an attributor is the wrong shape of statement rather than merely an imprecise one. For the 44 molecular rows in the matrix, correctness cannot be measured at all with the labels the field currently has, which is the gap that most needs closing. The audit is what makes those failures visible; the pipeline, the figures and every number here regenerate from the committed artifacts as the grid fills in.

9 Reproducibility and compute

What is pinned. Every result in this paper comes from one invocation of `full.yaml`, seeds 0, 1, 2, launched at git revision `aad466a` — the revision the manifest records, and the state of the pipeline when the sweep started; its outputs were committed afterwards, so that revision is not itself a reproducible checkout. The code, configurations and results together are archived at tag `v1.0.0-jcheminf`, which is the version of record for this manuscript and the state a reader should obtain. The run was on Python 3.12.13 with torch 2.11.0+cu128, PyTorch Geometric 2.8.0.post1, RDKit 2026.03.5, Captum 0.9.0, NumPy 2.0.2 and scikit-learn 1.6.1. The run manifest records all of these, the seed, the platform string and the CUDA device count (1 device visible), and is committed alongside the results. Seeds are set for Python, NumPy and torch before every stage; splits are deterministic functions of the dataset and seed, not of iteration order.

How to reproduce. The sweep is a single unattended entry point, `python -m molsanity.run_all --config configs/full.yaml`, with a minutes-long `configs/smoke.yaml` for a smoke test; neither prompts for input. `make -C paper` then regenerates every figure, table and quoted statistic in this manuscript from the results directory and fails the build if the prose quotes

a macro the current results do not define. Datasets download on first use and are checksummed against the committed manifest. No step requires credentials, and the credential-gated resources named in Section Experimental setup are out of scope and were never accessed.

Resumability and caching. Each cell-run writes a marker keyed to a content hash of its own configuration — cell, split, budget, model, training schedule and seed — so re-invoking the sweep re-executes exactly the cells whose configuration changed and reuses the rest. The hash is also what makes a correction auditable: when we found and fixed a defect in the scaffold split (§5.8), bumping a split-version field in the hash invalidated the affected cells and left the random-split cells untouched, and the re-run changed 0 of the random rows. Trained checkpoints are content-addressed on the same principle and, since the fix, on a digest of the realised train/val/test partition rather than on the split's name, so a checkpoint cannot silently outlive the partition it was trained for.

Device policy. Training uses the GPU; the attribution loop is CPU-only, because the worker pool forks and a CUDA context does not survive `fork`. This is a deliberate constraint, and it has a reproducibility consequence worth stating: we verified that Integrated Gradients, Saliency, Guided Backpropagation and Input×Gradient are bit-identical on GPU and CPU, whereas GNNExplainer moves by 0.003–0.012 because it fits a mask with Adam over ∼100 steps. A reader reproducing on a GPU should expect to match GNNExplainer to two decimals, not more. We also verified that the parallel audit loop reproduces the serial one field-by-field over complete per-molecule records, so cores may be traded for wall time freely.

One defect here was real and is worth recording. DIG's SubgraphX draws from `np.random`, which `torch.manual_seed` never resets, so its output depended on how many molecules had been attributed before it — results that changed if a run resumed at a different point. Seeding per molecule from (run seed, graph id) makes each attribution a function of the molecule alone, which is both order-independent and resume-safe.

Cost. The committed log covers one invocation, 2026-08-05 19:14 to 2026-08-06 09:17: 14.1 hours of wall clock over 558 cell-runs, of which 224 were executed and 334 served from cache. Because the sweep resumed onto a partly populated cache, that figure is *not* what a cold run costs, and we do not present it as one.

For a reader starting from an empty cache we give an estimate rather than a measurement, and label it as such: pricing every one of the 558 planned cell-runs at the median observed in this log for its own attributor gives roughly 32 hours on a single GPU. The audit is embarrassingly parallel across cells and resumable at cell granularity, so that is a serial upper bound rather than a required sitting.

The per-unit costs are what generalise. Median training time for a (dataset, backbone, split, seed) checkpoint is 26 s; training is negligible next to attribution. Attribution cost is concentrated rather than spread: the median SubgraphX cell-run takes 22 minutes against 2 for Integrated Gradients, a factor of 11 at equal sample size, and the 15 SubgraphX cell-runs alone account for 55% of the 14.0 hours of attribution compute in this log.

We report that ratio as a finding rather than as housekeeping. Compute cost is a selection pressure on which attributors get benchmarked at all, and it pushes the literature toward cheap ones; a benchmark that silently caps the expensive attributor's n to fit a budget buys its coverage with exactly the small-sample weakness the field criticises. SubgraphX is therefore held at the full per-cell sample and the number of cells reduced instead, which makes the coverage limit explicit (§7) rather than hiding it inside a confidence interval.

10 Acknowledgements

This work was carried out at La Trobe University. It builds directly on open-source software and released benchmarks: PyTorch Geometric [18], Captum [17], RDKit [19], DIG [12] and GraphXAI [10], and the MUTAG [29], MoleculeNet [27] and Therapeutics Data Commons [28] collections. MolSanity wraps the canonical implementations of the attribution methods it audits rather than reimplementing them, and would not be possible without the authors of those packages releasing them openly. Compute for the reported sweep was provided by Google Colab.

Generative AI disclosure. An AI coding assistant was used during software development and manuscript preparation. It was not used to generate, select or interpret results: every number, figure and table is computed by the released code from the committed artifacts (§9). The authors designed the study, verified the analyses and take full responsibility for the content.

11 Data and code availability

All code, configurations, trained checkpoints, per-combination results, per-molecule audit records and the run manifest (seed, library versions, hardware, and the per-combination outcome ledger) are available at

https://github.com/Kar488/molsanity, released under the MIT licence. The state corresponding to this manuscript is tag `v1.0.0-jcheminf`, from which a versioned archive including the trained weights is deposited to Zenodo; the repository carries `CITATION.cff` and `.zenodo.json` so that archive is generated from the tagged release rather than assembled by hand. Continuous integration rebuilds every figure, table and quoted statistic in this manuscript from the committed results on each change, so a divergence between the two fails the build.

Additional files.

- **Additional file 1** — `Additional_file_1.pdf`. Supplementary notes and summary tables: the ground-truth benchmark summary per dataset and split, the rationale-use partition testing the Faber objection, and the abstention coverage–reliability analysis.
- **Additional file 2** — `Additional_file_2.csv`. Per-cell audit matrix, one row per (dataset, backbone, attributor, split, seed) combination; the table every figure and quoted statistic in this manuscript is computed from.
- **Additional file 3** — `Additional_file_3.csv`. Across-seed mean, standard deviation, minimum and maximum per cell over the 3 seeds.
- **Additional file 4** — `Additional_file_4.csv`. Head-to-head benchmark: every audited cell scored on the MolSanity axes alongside Fidelity±, sparsity, the GraphFramEx characterisation score and the PyG/DIG unfaithfulness metric, computed on the same molecules.

All four are generated from the committed artifacts by `paper/make_additional_files.py`, so they cannot describe a different run from the manuscript.

Every figure and table in this paper is generated from those artifacts by `paper/figs/make_figures.py` and `paper/figs/make_tables.py`, which read the released results directly; no value in this manuscript is transcribed by hand. Re-running them after a new sweep updates every number, figure and table in place. Dataset licences and provenance are recorded in the repository's dataset manifest; the credential-gated resources referenced in Section Experimental setup are out of scope and were never accessed.

Table 4: Paired attributor comparisons on shared molecules, computed from the committed per-molecule records (Wilcoxon signed-rank on per-molecule occlusion faithfulness, $\Delta = A - B$). Same cell family as Table 5. p is the raw two-sided Wilcoxon value and q its Benjamini–Hochberg adjustment, controlling the false discovery rate over the contrasts within each cell block, which is the family a reader compares across.

A vs. B	n	median Δ	p	q
MUTAG, GINE, random split				
IG vs. Saliency	58	0.000	0.885	0.885
IG vs. Input×Grad	58	0.000	0.862	0.885
IG vs. GuidedBP	58	0.000	0.046	0.108
IG vs. GNNExpl.	58	0.000	0.704	0.861
IG vs. PGExpl.	57	0.149	0.006	0.019
IG vs. SubgraphX	57	0.000	0.692	0.861
Saliency vs. Input×Grad	58	0.000	0.779	0.861
Saliency vs. GuidedBP	58	0.000	0.005	0.018
Saliency vs. GNNExpl.	58	0.000	0.597	0.861
Saliency vs. PGExpl.	57	0.193	<0.001	0.001
Saliency vs. SubgraphX	57	0.000	0.689	0.861
Input×Grad vs. GuidedBP	58	0.000	0.005	0.018
Input×Grad vs. GNNExpl.	58	0.000	0.608	0.861
Input×Grad vs. PGExpl.	57	0.213	<0.001	0.001
Input×Grad vs. SubgraphX	57	−0.015	0.711	0.861
GuidedBP vs. GNNExpl.	58	0.046	0.148	0.294
GuidedBP vs. PGExpl.	57	0.316	<0.001	<0.001
GuidedBP vs. SubgraphX	57	0.028	0.154	0.294
GNNExpl. vs. PGExpl.	57	0.070	0.007	0.019
GNNExpl. vs. SubgraphX	57	0.000	0.774	0.861
PGExpl. vs. SubgraphX	56	−0.167	<0.001	0.001
MUTAG, GINE, scaffold split				
IG vs. Saliency	53	0.000	0.098	0.158
IG vs. Input×Grad	53	−0.200	0.002	0.005
IG vs. GuidedBP	53	−0.357	<0.001	<0.001
IG vs. GNNExpl.	53	0.000	0.276	0.369
IG vs. PGExpl.	52	0.000	0.525	0.613
IG vs. SubgraphX	53	0.000	0.355	0.438
Saliency vs. Input×Grad	53	0.000	0.029	0.060
Saliency vs. GuidedBP	53	−0.107	<0.001	<0.001
Saliency vs. GNNExpl.	53	0.000	0.722	0.746
Saliency vs. PGExpl.	52	0.081	0.044	0.085
Saliency vs. SubgraphX	53	0.033	0.660	0.729
Input×Grad vs. GuidedBP	53	−0.058	<0.001	0.001
Input×Grad vs. GNNExpl.	53	0.017	0.019	0.044
Input×Grad vs. PGExpl.	52	0.207	<0.001	0.001
Input×Grad vs. SubgraphX	53	0.023	0.137	0.206
GuidedBP vs. GNNExpl.	53	0.200	<0.001	<0.001
GuidedBP vs. PGExpl.	52	0.300	<0.001	<0.001
GuidedBP vs. SubgraphX	53	0.200	<0.001	<0.001
GNNExpl. vs. PGExpl.	52	0.081	0.082	0.144
GNNExpl. vs. SubgraphX	53	0.000	0.746	0.746
PGExpl. vs. SubgraphX	52	0.000	0.281	0.369

Table 5: Does a faithfulness-only ranking pick the ground-truth-best attributor? 5 molecular cell families, each audited on both splits with the same backbone and the same 7 attributors, so within a family only the split changes and the contrast isolates distribution shift rather than confounding it with a change of dataset. Each row ranks the attributors by one faithfulness/fidelity metric and asks whether its top choice is the one the ground truth ranks best. p is a paired Wilcoxon test on per-molecule GT AUROC between the selected and the ground-truth-best attributor, over the molecules both audited, and q its Benjamini–Hochberg adjustment over all 30 selection tests in this table. Because a paired per-molecule test needs the same molecules on both sides and the split is a function of the seed, this table is computed within a single seed rather than across the three; the across-seed spread for the same cells is supplementary material.

cell	regime	ranking metric	its top pick	pick GT	GT-best	GT-best	mismatch	Wilcoxon p	q	ρ(faith,GT)
MUTAG·GINE, n=58	in-distribution	occlusion ρ	GuidedBP	0.037	GNNExpl.	0.858	**yes**	<0.001	<0.001	0.143
		Fidelity+	IG	0.496	GNNExpl.	0.858	**yes**	<0.001	<0.001	−0.357
		characterisation	SubgraphX	0.348	GNNExpl.	0.858	**yes**	<0.001	<0.001	0.821
MUTAG·GINE, n=53	scaffold shift	occlusion ρ	GuidedBP	0.013	GNNExpl.	0.826	**yes**	<0.001	<0.001	−0.643
		Fidelity+	SubgraphX	0.330	GNNExpl.	0.826	**yes**	<0.001	<0.001	−0.321
		characterisation	SubgraphX	0.330	GNNExpl.	0.826	**yes**	<0.001	<0.001	0.036
MolMotifHard·GINE, n=200	in-distribution	occlusion ρ	SubgraphX	0.421	Saliency	0.994	**yes**	<0.001	<0.001	−0.286
		Fidelity+	SubgraphX	0.421	Saliency	0.994	**yes**	<0.001	<0.001	0.107
		characterisation	SubgraphX	0.421	Saliency	0.994	**yes**	<0.001	<0.001	−0.357
MolMotifHard·GINE, n=200	scaffold shift	occlusion ρ	PGExpl.	0.565	IG	1.000	**yes**	<0.001	<0.001	−0.893
		Fidelity+	IG	1.000	IG	1.000	no	—	—	0.821
		characterisation	IG	1.000	IG	1.000	no	—	—	0.857
MolMotif·GINE, n=200	in-distribution	occlusion ρ	GNNExpl.	0.630	Input×Grad	0.999	**yes**	<0.001	<0.001	−0.571
		Fidelity+	GuidedBP	0.998	Input×Grad	0.999	**yes**	0.091	0.091	0.750
		characterisation	SubgraphX	0.556	Input×Grad	0.999	**yes**	<0.001	<0.001	−0.107
MolMotif·GINE, n=200	scaffold shift	occlusion ρ	IG	0.939	Input×Grad	0.998	**yes**	<0.001	<0.001	−0.071
		Fidelity+	IG	0.939	Input×Grad	0.998	**yes**	<0.001	<0.001	0.036
		characterisation	IG	0.939	Input×Grad	0.998	**yes**	<0.001	<0.001	0.143
Benzene·GINE, n=200	in-distribution	occlusion ρ	GuidedBP	0.743	Input×Grad	1.000	**yes**	<0.001	<0.001	−0.054
		Fidelity+	SubgraphX	0.457	Input×Grad	1.000	**yes**	<0.001	<0.001	−0.432
		characterisation	SubgraphX	0.457	Input×Grad	1.000	**yes**	<0.001	<0.001	0.126
Benzene·GINE, n=200	scaffold shift	occlusion ρ	GuidedBP	0.997	IG	0.999	**yes**	0.043	0.045	0.607
		Fidelity+	GuidedBP	0.997	IG	0.999	**yes**	0.043	0.045	0.893
		characterisation	SubgraphX	0.770	IG	0.999	**yes**	<0.001	<0.001	0.571
FluorideCarbonyl·GINE, n=200	in-distribution	occlusion ρ	GuidedBP	0.881	GuidedBP	0.881	no	—	—	0.786
		Fidelity+	SubgraphX	0.538	GuidedBP	0.881	**yes**	<0.001	<0.001	0.000
		characterisation	SubgraphX	0.538	GuidedBP	0.881	**yes**	<0.001	<0.001	0.250
FluorideCarbonyl·GINE, n=200	scaffold shift	occlusion ρ	GuidedBP	0.937	GuidedBP	0.937	no	—	—	0.750
		Fidelity+	IG	0.833	GuidedBP	0.937	**yes**	<0.001	<0.001	0.893
		characterisation	IG	0.833	GuidedBP	0.937	**yes**	<0.001	<0.001	0.893

Table 6: Every committed cell in which attribution *correctness* is measurable. Ground truth is exact for the synthetic sets and for MolMotif/MolMotifHard (the label *is* the substructure), and a chemically motivated nitro-motif proxy for MUTAG. GT AUROC $= 0.5$ is chance; below 0.5 the attribution is anti-aligned with the true motif. Occlusion ρ is the attribution–occlusion rank agreement (higher = more faithful to the model). Every row was produced by the single run in Table 2; none are carried over from an earlier one. Bold marks the best GT AUROC and the best occlusion ρ within each dataset×split block (emphasis only; values are unaltered).

dataset	backbone	attributor	split	n	acc	AUC	GT AUROC	GT AUPRC	occ. ρ	Fid+	Fid−	stab.
BA-2Motifs	GINE	GNNExpl.	random	200	0.767	1.000	0.511	0.308	−0.027	−0.001	0.019	0.951
BA-2Motifs	GINE	GuidedBP	random	200	0.767	1.000	0.945	0.901	−0.248	0.039	−0.006	0.745
BA-2Motifs	GINE	Input×Grad	random	200	0.767	1.000	0.975	0.908	−0.191	0.030	0.001	0.796
BA-2Motifs	GINE	IG	random	200	0.767	1.000	**0.997**	0.988	−0.204	0.030	0.001	0.645
BA-2Motifs	GINE	PGExpl.	random	200	0.767	1.000	0.398	0.304	**0.015**	0.006	0.027	0.305
BA-2Motifs	GINE	Saliency	random	200	0.767	1.000	0.975	0.908	−0.191	0.030	0.001	0.796
BA-2Motifs	GINE	GNNExpl.	scaffold	200	0.603	0.978	0.563	0.335	−0.041	−0.002	−0.007	0.969
BA-2Motifs	GINE	GuidedBP	scaffold	200	0.603	0.978	0.843	0.618	−0.253	0.000	−0.004	0.807
BA-2Motifs	GINE	Input×Grad	scaffold	200	0.603	0.978	**0.937**	0.789	−0.286	−0.001	−0.005	0.663
BA-2Motifs	GINE	IG	scaffold	200	0.603	0.978	0.856	0.543	−0.310	−0.003	−0.004	0.715
BA-2Motifs	GINE	PGExpl.	scaffold	200	0.603	0.978	0.882	0.763	**0.032**	−0.002	−0.005	0.092
BA-2Motifs	GINE	Saliency	scaffold	200	0.603	0.978	0.937	0.789	−0.286	−0.001	−0.005	0.663
Benzene	GINE	GNNExpl.	random	200	0.872	1.000	0.512	0.441	0.004	0.136	0.293	0.733
Benzene	GINE	GuidedBP	random	200	0.872	1.000	0.897	0.853	**0.384**	0.267	0.198	0.560
Benzene	GINE	Input×Grad	random	200	0.872	1.000	0.998	0.996	0.205	0.250	0.225	0.566
Benzene	GINE	IG	random	200	0.872	1.000	**1.000**	1.000	0.374	0.251	0.200	0.525
Benzene	GINE	PGExpl.	random	200	0.872	1.000	0.315	0.422	−0.311	0.037	0.284	0.539
Benzene	GINE	Saliency	random	200	0.872	1.000	0.997	0.995	0.152	0.248	0.244	0.629
Benzene	GINE	SubgraphX	random	200	0.872	1.000	0.503	0.451	0.156	0.261	0.063	—
Benzene	GINE	GNNExpl.	scaffold	200	0.925	1.000	0.549	0.499	0.270	0.161	0.257	0.986
Benzene	GINE	GuidedBP	scaffold	200	0.925	1.000	0.881	0.877	**0.351**	0.257	0.194	0.804
Benzene	GINE	Input×Grad	scaffold	200	0.925	1.000	0.995	0.989	0.181	0.226	0.220	0.865
Benzene	GINE	IG	scaffold	200	0.925	1.000	**0.998**	0.997	0.279	0.232	0.210	0.841
Benzene	GINE	PGExpl.	scaffold	200	0.925	1.000	0.259	0.365	−0.146	0.029	0.244	0.366
Benzene	GINE	Saliency	scaffold	200	0.925	1.000	0.994	0.988	0.208	0.219	0.213	0.905
Benzene	GINE	SubgraphX	scaffold	200	0.925	1.000	0.592	0.514	0.220	0.167	0.101	—
FluorideCarbonyl	GINE	GNNExpl.	random	200	0.945	0.974	0.600	0.513	0.086	0.153	0.291	0.460
FluorideCarbonyl	GINE	GuidedBP	random	200	0.945	0.974	**0.884**	0.789	**0.170**	0.148	0.290	0.289
FluorideCarbonyl	GINE	Input×Grad	random	200	0.945	0.974	0.812	0.722	0.045	0.270	0.181	0.344
FluorideCarbonyl	GINE	IG	random	200	0.945	0.974	0.758	0.696	0.132	0.333	0.135	0.410
FluorideCarbonyl	GINE	PGExpl.	random	200	0.945	0.974	0.239	0.298	−0.119	0.080	0.338	−0.101
FluorideCarbonyl	GINE	Saliency	random	200	0.945	0.974	0.852	0.752	0.080	0.203	0.229	0.450
FluorideCarbonyl	GINE	SubgraphX	random	200	0.945	0.974	0.491	0.379	−0.001	0.196	0.254	—
FluorideCarbonyl	GINE	GNNExpl.	scaffold	200	0.957	0.985	0.631	0.526	−0.078	0.155	0.122	0.911
FluorideCarbonyl	GINE	GuidedBP	scaffold	200	0.957	0.985	**0.946**	0.900	**0.185**	0.222	0.090	0.660
FluorideCarbonyl	GINE	Input×Grad	scaffold	200	0.957	0.985	0.777	0.686	0.152	0.192	0.118	0.595
FluorideCarbonyl	GINE	IG	scaffold	200	0.957	0.985	0.797	0.732	0.130	0.274	0.085	0.395
FluorideCarbonyl	GINE	PGExpl.	scaffold	200	0.957	0.985	0.346	0.321	−0.067	0.037	0.240	0.356
FluorideCarbonyl	GINE	Saliency	scaffold	200	0.957	0.985	0.876	0.774	0.150	0.189	0.104	0.706
FluorideCarbonyl	GINE	SubgraphX	scaffold	200	0.957	0.985	0.467	0.323	0.016	0.187	0.103	—

Table 7: Ground-truth cells, continued (2 of 4). Columns and conventions as in Table 6.

dataset	backbone	attributor	split	n	acc	AUC	GT AUROC	GT AUPRC	occ. ρ	Fid+	Fid−	stab.
MUTAG	AttentiveFP	IG	random	58	0.735	0.915	0.040	0.171	0.156	0.065	0.124	0.822
MUTAG	GAT	IG	random	58	0.621	0.672	0.400	0.379	0.151	0.209	0.296	0.951
MUTAG	GCN	IG	random	58	0.638	0.700	0.367	0.368	**0.496**	0.237	0.302	0.911
MUTAG	GINE	GNNExpl.	random	58	0.747	0.912	**0.678**	0.590	−0.121	0.189	0.205	0.848
MUTAG	GINE	GuidedBP	random	58	0.747	0.912	0.097	0.197	−0.131	0.330	0.351	0.882
MUTAG	GINE	Input×Grad	random	58	0.747	0.912	0.035	0.169	−0.172	0.295	0.354	0.770
MUTAG	GINE	IG	random	58	0.747	0.912	0.387	0.356	−0.177	0.330	0.313	0.724
MUTAG	GINE	PGExpl.	random	58	0.747	0.912	0.663	0.600	−0.159	0.246	0.292	0.713
MUTAG	GINE	Saliency	random	58	0.747	0.912	0.012	0.165	−0.151	0.294	0.356	0.767
MUTAG	GINE	SubgraphX	random	58	0.747	0.912	0.393	0.308	−0.035	0.276	0.029	—
MUTAG	MPNN	IG	random	58	0.787	0.722	0.103	0.181	0.358	0.506	0.516	0.890
MUTAG	AttentiveFP	IG	scaffold	53	0.774	0.886	0.029	0.134	0.325	0.008	0.017	0.907
MUTAG	GAT	IG	scaffold	53	0.824	0.913	0.533	0.506	0.607	0.054	0.171	1.000
MUTAG	GCN	IG	scaffold	53	0.799	0.894	0.083	0.161	**0.763**	0.159	0.340	0.941
MUTAG	GINE	GNNExpl.	scaffold	53	0.736	0.808	**0.774**	0.695	0.124	0.116	0.145	0.904
MUTAG	GINE	GuidedBP	scaffold	53	0.736	0.808	0.047	0.143	0.337	0.268	0.128	0.930
MUTAG	GINE	Input×Grad	scaffold	53	0.736	0.808	0.116	0.154	0.247	0.229	0.196	0.826
MUTAG	GINE	IG	scaffold	53	0.736	0.808	0.568	0.449	0.163	0.159	0.227	0.808
MUTAG	GINE	PGExpl.	scaffold	53	0.736	0.808	0.319	0.205	0.056	0.038	0.202	0.606
MUTAG	GINE	Saliency	scaffold	53	0.736	0.808	0.051	0.143	0.216	0.227	0.174	0.854
MUTAG	GINE	SubgraphX	scaffold	53	0.736	0.808	0.392	0.279	0.200	0.240	0.023	—
MUTAG	MPNN	IG	scaffold	53	0.837	0.913	0.163	0.185	0.206	0.166	0.253	0.829
MolMotif	AttentiveFP	IG	random	200	0.955	0.996	0.967	0.923	**0.496**	0.469	0.025	0.912
MolMotif	GAT	IG	random	200	1.000	1.000	0.842	0.776	0.338	0.440	0.225	0.901
MolMotif	GCN	IG	random	200	0.872	0.996	0.943	0.895	0.167	0.324	0.497	1.000
MolMotif	GINE	GNNExpl.	random	200	0.928	1.000	0.641	0.437	−0.017	0.184	0.296	0.973
MolMotif	GINE	GuidedBP	random	200	0.928	1.000	0.971	0.928	−0.113	0.205	0.293	0.884
MolMotif	GINE	Input×Grad	random	200	0.928	1.000	**0.995**	0.989	−0.179	0.146	0.300	0.848
MolMotif	GINE	IG	random	200	0.928	1.000	0.951	0.875	−0.004	0.341	0.218	0.970
MolMotif	GINE	PGExpl.	random	200	0.928	1.000	0.502	0.335	−0.098	0.132	0.302	0.639
MolMotif	GINE	Saliency	random	200	0.928	1.000	0.994	0.987	−0.164	0.161	0.300	0.837
MolMotif	GINE	SubgraphX	random	200	0.928	1.000	0.555	0.327	−0.036	0.215	0.215	—
MolMotif	MPNN	IG	random	200	1.000	1.000	0.873	0.729	−0.124	0.377	0.347	0.946
MolMotif	AttentiveFP	IG	scaffold	200	0.985	1.000	0.923	0.870	**0.672**	0.447	0.024	0.929
MolMotif	GAT	IG	scaffold	200	0.995	1.000	0.798	0.723	0.332	0.449	0.454	0.945
MolMotif	GCN	IG	scaffold	200	0.970	1.000	0.988	0.975	0.391	0.399	0.177	0.933
MolMotif	GINE	GNNExpl.	scaffold	200	0.973	1.000	0.616	0.425	0.054	0.301	0.344	0.982
MolMotif	GINE	GuidedBP	scaffold	200	0.973	1.000	0.989	0.977	0.033	0.284	0.346	0.843
MolMotif	GINE	Input×Grad	scaffold	200	0.973	1.000	**0.999**	0.997	−0.086	0.156	0.360	0.869
MolMotif	GINE	IG	scaffold	200	0.973	1.000	0.897	0.749	0.058	0.376	0.313	0.880
MolMotif	GINE	PGExpl.	scaffold	200	0.973	1.000	0.513	0.361	−0.111	0.129	0.299	0.707
MolMotif	GINE	Saliency	scaffold	200	0.973	1.000	0.997	0.991	−0.090	0.139	0.367	0.866
MolMotif	GINE	SubgraphX	scaffold	200	0.973	1.000	0.469	0.265	−0.007	0.362	0.129	—
MolMotif	MPNN	IG	scaffold	200	0.998	1.000	0.871	0.727	0.055	0.399	0.348	0.922

Table 8: Ground-truth cells, continued (3 of 4). Columns and conventions as in Table 6.

dataset	backbone	attributor	split	n	acc	AUC	GT AUROC	GT AUPRC	occ. ρ	Fid+	Fid−	stab.
MolMotifHard	AttentiveFP	IG	random	200	0.972	0.981	0.878	0.823	**0.476**	0.411	0.087	0.932
MolMotifHard	GAT	IG	random	200	0.962	0.992	0.793	0.770	0.041	0.466	0.404	0.966
MolMotifHard	GCN	IG	random	200	0.885	0.988	0.901	0.855	0.040	0.359	0.355	0.700
MolMotifHard	GINE	GNNExpl.	random	200	0.935	0.998	0.413	0.385	−0.075	0.141	0.251	0.942
MolMotifHard	GINE	GuidedBP	random	200	0.935	0.998	0.919	0.890	−0.050	0.204	0.243	0.884
MolMotifHard	GINE	Input×Grad	random	200	0.935	0.998	0.996	0.994	−0.068	0.228	0.221	0.925
MolMotifHard	GINE	IG	random	200	0.935	0.998	0.985	0.973	−0.074	0.218	0.231	0.914
MolMotifHard	GINE	PGExpl.	random	200	0.935	0.998	0.512	0.419	−0.051	0.162	0.208	0.830
MolMotifHard	GINE	Saliency	random	200	0.935	0.998	**0.998**	0.996	−0.064	0.217	0.217	0.911
MolMotifHard	GINE	SubgraphX	random	200	0.935	0.998	0.408	0.320	−0.040	0.201	0.150	—
MolMotifHard	MPNN	IG	random	200	0.908	0.996	0.953	0.920	0.310	0.289	0.222	0.933
MolMotifHard	AttentiveFP	IG	scaffold	200	0.975	0.995	0.900	0.825	0.316	0.388	0.075	0.893
MolMotifHard	GAT	IG	scaffold	200	0.958	0.997	0.807	0.737	−0.045	0.382	0.356	0.921
MolMotifHard	GCN	IG	scaffold	200	0.845	0.995	0.866	0.761	0.026	0.261	0.401	0.637
MolMotifHard	GINE	GNNExpl.	scaffold	200	0.892	0.997	0.420	0.349	−0.167	0.180	0.225	0.960
MolMotifHard	GINE	GuidedBP	scaffold	200	0.892	0.997	0.875	0.818	−0.136	0.285	0.186	0.794
MolMotifHard	GINE	Input×Grad	scaffold	200	0.892	0.997	0.995	0.990	−0.173	0.273	0.199	0.873
MolMotifHard	GINE	IG	scaffold	200	0.892	0.997	0.994	0.986	−0.201	0.308	0.188	0.909
MolMotifHard	GINE	PGExpl.	scaffold	200	0.892	0.997	0.376	0.300	−0.085	0.087	0.234	0.717
MolMotifHard	GINE	Saliency	scaffold	200	0.892	0.997	**0.999**	0.995	−0.165	0.278	0.211	0.847
MolMotifHard	GINE	SubgraphX	scaffold	200	0.892	0.997	0.472	0.310	−0.130	0.238	0.175	—
MolMotifHard	MPNN	IG	scaffold	200	0.987	0.998	0.965	0.924	**0.578**	0.480	0.093	0.886
ShapeGGen	GINE	GNNExpl.	random	50	0.780	0.578	0.549	0.591	0.389	0.239	0.155	0.563
ShapeGGen	GINE	GuidedBP	random	50	0.780	0.578	**0.773**	0.811	0.481	0.239	0.161	0.796
ShapeGGen	GINE	Input×Grad	random	50	0.780	0.578	0.725	0.764	0.435	0.250	0.171	0.744
ShapeGGen	GINE	IG	random	50	0.780	0.578	0.733	0.771	0.447	0.266	0.154	0.763
ShapeGGen	GINE	PGExpl.	random	50	0.780	0.578	0.499	0.523	−0.163	0.127	0.299	0.676
ShapeGGen	GINE	Saliency	random	50	0.780	0.578	0.769	0.808	0.400	0.228	0.191	0.730
ShapeGGen	GINE	SubgraphX	random	50	0.780	0.578	0.644	0.592	**0.543**	0.403	−0.010	0.233
ShapeGGen	GINE	GNNExpl.	scaffold	50	0.580	0.552	0.574	0.628	0.143	0.028	0.020	0.417
ShapeGGen	GINE	GuidedBP	scaffold	50	0.580	0.552	**0.803**	0.826	0.227	0.027	0.015	0.593
ShapeGGen	GINE	Input×Grad	scaffold	50	0.580	0.552	0.708	0.735	0.143	0.021	0.022	0.494
ShapeGGen	GINE	IG	scaffold	50	0.580	0.552	0.727	0.749	0.111	0.025	0.019	0.505
ShapeGGen	GINE	PGExpl.	scaffold	50	0.580	0.552	0.567	0.611	−0.062	0.021	0.034	0.653
ShapeGGen	GINE	Saliency	scaffold	50	0.580	0.552	0.776	0.801	0.164	0.020	0.025	0.445
ShapeGGen	GINE	SubgraphX	scaffold	50	0.580	0.552	0.598	0.580	**0.407**	0.059	−0.006	0.217

Table 9: Ground-truth cells, continued (4 of 4). Columns and conventions as in Table 6.

dataset	backbone	attributor	split	n	acc	AUC	GT AUROC	GT AUPRC	occ. ρ	Fid+	Fid−	stab.
SynthMotifs	AttentiveFP	IG	random	200	0.980	0.999	0.871	0.721	0.185	0.441	0.230	0.726
SynthMotifs	GAT	IG	random	200	0.990	1.000	0.793	0.660	0.242	0.284	0.209	0.384
SynthMotifs	GCN	IG	random	200	0.888	1.000	0.852	0.767	0.415	0.323	0.125	0.694
SynthMotifs	GINE	GNNExpl.	random	200	0.972	1.000	0.632	0.388	0.123	0.232	0.251	0.698
SynthMotifs	GINE	GuidedBP	random	200	0.972	1.000	**0.995**	0.983	**0.640**	0.272	0.061	0.795
SynthMotifs	GINE	Input×Grad	random	200	0.972	1.000	0.970	0.899	0.475	0.273	0.111	0.656
SynthMotifs	GINE	IG	random	200	0.972	1.000	0.990	0.960	0.472	0.265	0.095	0.849
SynthMotifs	GINE	PGExpl.	random	200	0.972	1.000	0.415	0.209	−0.084	0.091	0.226	0.427
SynthMotifs	GINE	Saliency	random	200	0.972	1.000	0.979	0.927	0.489	0.274	0.105	0.560
SynthMotifs	GINE	SubgraphX	random	200	0.972	1.000	0.770	0.533	0.459	0.245	0.056	0.744
SynthMotifs	MPNN	IG	random	200	0.993	1.000	0.872	0.694	0.337	0.423	0.417	0.590
SynthMotifs	AttentiveFP	IG	scaffold	200	0.963	1.000	0.870	0.689	0.270	0.374	0.203	0.741
SynthMotifs	GAT	IG	scaffold	200	0.872	1.000	0.912	0.820	0.431	0.496	0.269	1.000
SynthMotifs	GCN	IG	scaffold	200	0.860	0.996	0.953	0.825	0.352	0.246	0.144	0.625
SynthMotifs	GINE	GNNExpl.	scaffold	200	0.947	1.000	0.642	0.397	0.092	0.327	0.341	0.637
SynthMotifs	GINE	GuidedBP	scaffold	200	0.947	1.000	**0.974**	0.928	**0.623**	0.424	0.045	0.732
SynthMotifs	GINE	Input×Grad	scaffold	200	0.947	1.000	0.955	0.866	0.421	0.403	0.133	0.705
SynthMotifs	GINE	IG	scaffold	200	0.947	1.000	0.973	0.901	0.459	0.412	0.065	0.741
SynthMotifs	GINE	PGExpl.	scaffold	200	0.947	1.000	0.354	0.181	−0.073	0.105	0.339	0.585
SynthMotifs	GINE	Saliency	scaffold	200	0.947	1.000	0.957	0.886	0.401	0.394	0.150	0.601
SynthMotifs	GINE	SubgraphX	scaffold	200	0.947	1.000	0.751	0.499	0.284	0.377	0.045	0.562
SynthMotifs	MPNN	IG	scaffold	200	0.838	0.999	0.845	0.652	0.225	0.317	0.288	0.550

Table 10: Molecular classification cells (no node-level ground truth available), 1 of 2. Every committed classification cell on a real molecular dataset. The ground-truth localisation column does not exist for these datasets, since no per-atom labels are published, so only model-side reliability can be measured, which is the gap the Tier-1 cells are needed to close. *charact.* is the GraphFramEx characterisation score and *unfaith.* the PyG/DIG unfaithfulness metric, computed on the same molecules. Bold marks the highest occlusion ρ per dataset (emphasis only).

dataset	backbone	attributor	split	n	acc	AUC	ECE	motif top-1	occ. ρ	occ. top-1	Fid+	Fid−	stab.	charact.	unfaith.
BACE	GCN	IG	random	200	0.720	0.819	0.052	0.771	**0.235**	0.633	0.344	0.366	0.849	0.291	0.594
BACE	GCN	IG	scaffold	200	0.628	0.698	0.116	0.807	0.138	0.567	0.352	0.352	0.865	0.227	0.668
BACE	GINE	IG	random	200	0.730	0.821	0.062	0.788	0.116	0.557	0.274	0.268	0.691	0.126	0.323
BACE	GINE	IG	scaffold	200	0.545	0.716	0.061	0.863	0.093	0.297	0.239	0.271	0.564	0.415	0.174
BBBP	AttentiveFP	IG	random	200	0.757	0.884	0.038	0.749	0.234	0.418	0.092	0.093	0.853	0.200	0.104
BBBP	AttentiveFP	IG	scaffold	200	0.855	0.954	0.127	0.750	0.409	0.562	0.251	0.169	0.847	0.327	0.070
BBBP	GAT	IG	random	200	0.735	0.900	0.032	0.770	0.002	0.447	0.219	0.219	0.876	0.193	0.291
BBBP	GAT	IG	scaffold	200	0.812	0.914	0.123	0.788	0.123	0.522	0.262	0.322	0.941	0.158	0.444
BBBP	GCN	IG	random	200	0.740	0.889	0.043	0.796	−0.143	0.425	0.155	0.157	0.842	0.092	0.195
BBBP	GCN	IG	scaffold	200	0.823	0.915	0.115	0.801	0.189	0.528	0.302	0.357	0.893	0.080	0.506
BBBP	GINE	GNNExpl.	random	200	0.730	0.892	0.033	0.799	−0.018	0.397	0.183	0.216	0.955	0.241	0.164
BBBP	GINE	GNNExpl.	scaffold	200	0.827	0.925	0.115	0.804	0.349	0.487	0.330	0.355	0.934	0.327	0.400
BBBP	GINE	IG	random	200	0.730	0.892	0.033	0.770	0.044	0.427	0.219	0.220	0.890	0.206	0.314
BBBP	GINE	IG	scaffold	200	0.827	0.925	0.115	0.763	**0.509**	0.500	0.282	0.363	0.855	0.277	0.471
BBBP	GINE	PGExpl.	random	200	0.730	0.892	0.033	0.899	−0.131	0.285	0.126	0.202	0.370	0.159	—
BBBP	GINE	PGExpl.	scaffold	200	0.827	0.925	0.115	0.882	0.212	0.258	0.189	0.331	0.438	0.223	—
BBBP	MPNN	IG	random	200	0.757	0.906	0.040	0.792	0.047	0.373	0.109	0.169	0.860	0.113	0.142
BBBP	MPNN	IG	scaffold	200	0.818	0.923	0.115	0.819	0.160	0.502	0.175	0.323	0.882	0.080	0.307

Table 11: Molecular classification cells (no node-level ground truth available), 2 of 2. Every committed classification cell on a real molecular dataset. The ground-truth localisation column does not exist for these datasets, since no per-atom labels are published, so only model-side reliability can be measured, which is the gap the Tier-1 cells are needed to close. *charact.* is the GraphFramEx characterisation score and *unfaith.* the PyG/DIG unfaithfulness metric, computed on the same molecules. Bold marks the highest occlusion ρ per dataset (emphasis only).

dataset	backbone	attributor	split	n	acc	AUC	ECE	motif top-1	occ. ρ	occ. top-1	Fid+	Fid−	stab.	charact.	unfaith.
ClinTox	GINE	GNNExpl.	random	200	0.730	0.880	0.101	0.476	**0.275**	0.433	0.215	0.174	0.975	0.129	0.340
ClinTox	GINE	GNNExpl.	scaffold	200	0.673	0.844	0.144	0.670	−0.027	0.437	0.237	0.222	0.988	0.129	0.319
ClinTox	GINE	IG	random	200	0.730	0.880	0.101	0.799	−0.287	0.390	0.177	0.174	0.911	0.129	0.205
ClinTox	GINE	IG	scaffold	200	0.673	0.844	0.144	0.756	−0.170	0.417	0.222	0.222	0.971	0.129	0.221
DILI	GINE	IG	random	142	0.714	0.761	0.117	0.742	0.299	0.472	0.192	0.209	0.813	0.220	0.287
DILI	GINE	IG	scaffold	142	0.737	0.801	0.085	0.826	**0.336**	0.582	0.237	0.261	0.921	0.282	0.527
SIDER	GCN	IG	random	200	0.643	0.660	0.061	0.744	0.159	0.568	0.100	0.120	0.750	0.154	0.054
SIDER	GCN	IG	scaffold	200	0.575	0.668	0.028	0.748	0.365	0.722	0.181	0.312	0.879	0.179	0.033
SIDER	GINE	IG	random	200	0.642	0.671	0.070	0.764	0.277	0.613	0.174	0.241	0.869	0.243	0.449
SIDER	GINE	IG	scaffold	200	0.555	0.627	0.033	0.753	**0.436**	0.655	0.233	0.287	0.866	0.293	0.511
Tox21	GINE	IG	random	200	0.940	0.820	0.113	0.749	**−0.269**	0.333	−0.048	−0.042	0.710	0.037	0.309
Tox21	GINE	IG	scaffold	200	0.942	0.745	0.040	0.794	−0.413	0.402	−0.070	−0.068	0.927	0.042	0.149
hERG	GINE	IG	random	197	0.655	0.821	0.109	0.761	0.240	0.624	0.384	0.384	0.953	0.188	0.809
hERG	GINE	IG	scaffold	197	0.472	0.653	0.468	0.791	**0.743**	0.877	0.730	0.730	0.943	0.074	0.642

References

[1] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning* (*ICML*), PMLR 70:3319–3328, 2017.

[2] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. *International Conference on Learning Representations* (*ICLR*) *Workshop*, 2014.

[3] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: the all convolutional net. *International Conference on Learning Representations* (*ICLR*) *Workshop*, 2015.

[4] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. GNNExplainer: generating explanations for graph neural networks. *Advances in Neural Information Processing Systems* (*NeurIPS*), 2019.

[5] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang. Parameterized explainer for graph neural network. *Advances in Neural Information Processing Systems* (*NeurIPS*), 2020.

[6] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji. On explainability of graph neural networks via subgraph explorations. *Proceedings of the 38th International Conference on Machine Learning* (*ICML*), PMLR 139:12241–12252, 2021.

[7] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann. Explainability methods for graph convolutional neural networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2019.

[8] H. Yuan, H. Yu, S. Gui, and S. Ji. Explainability in graph neural networks: a taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5782–5799, 2023.

[9] B. Sanchez-Lengeling, J. Wei, B. Lee, E. Reif, P. Wang, W. Qian, K. McCloskey, L. Colwell, and A. Wiltschko. Evaluating attribution for graph neural networks. *Advances in Neural Information Processing Systems* (*NeurIPS*), 2020.

[10] C. Agarwal, O. Queen, H. Lakkaraju, and M. Zitnik. Evaluating explainability for graph neural networks. *Scientific Data*, 10:144, 2023.

[11] K. Amara, Z. Ying, Z. Zhang, Z. Han, Y. Zhao, Y. Shan, U. Brandes, S. Schemm, and C. Zhang. GraphFramEx: towards systematic evaluation of explainability methods for graph neural networks. *Proceedings of the First Learning on Graphs Conference* (*LoG*), PMLR 198, 2022.

[12] M. Liu, Y. Luo, L. Wang, Y. Xie, H. Yuan, S. Gui, H. Yu, Z. Xu, J. Zhang, Y. Liu, K. Yan, H. Liu, C. Fu, B. M. Oztekin, X. Zhang, and S. Ji. DIG: a turnkey library for diving into graph deep learning research. *Journal of Machine Learning Research*, 22(240):1–9, 2021.

[13] M. Hiltscher, M. Bianciotto, and F. Grisoni. Explaining What Matters: Faithfulness in Molecular Deep Learning. *ChemRxiv*, 2026. doi:10.26434/chemrxiv.10001594/v1. Platform: `https://github.com/molML/MolFaith`.

[14] L. Faber, A. K. Moghaddam, and R. Wattenhofer. When comparing to ground truth is wrong: on evaluating GNN explanation methods. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (*KDD*), pages 332–341, 2021.

[15] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems* (*NeurIPS*), 2018.

[16] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems* (*NeurIPS*), 2019.

[17] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson. Captum: a unified and generic model interpretability library for PyTorch. *arXiv:2009.07896*, 2020.

[18] M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch Geometric. *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[19] G. Landrum et al. RDKit: open-source cheminformatics. `https://www.rdkit.org`.

[20] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? *International Conference on Learning Representations* (*ICLR*), 2019.

[21] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec. Strategies for pre-training graph neural networks. *International Conference on Learning Representations* (*ICLR*), 2020.

[22] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations* (*ICLR*), 2017.

[23] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. *International Conference on Learning Representations* (*ICLR*), 2018.

[24] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning* (*ICML*), PMLR 70:1263–1272, 2017.

[25] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, and M. Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, 2020.

[26] K. Iyer and N. Sabar. M-JEPA: predictive self-supervised learning for molecular graphs with scaffold-shift evaluation on Tox21. *Journal of Chemical Information and Modeling*, 66(14):8008–8021, 2026. doi:10.1021/acs.jcim.6c00828.

[27] 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. *Chemical Science*, 9(2):513–530, 2018.

[28] K. Huang, T. Fu, W. Gao, Y. Zhao, Y. Roohani, J. Leskovec, C. W. Coley, C. Xiao, J. Sun, and M. Zitnik. Therapeutics Data Commons: machine learning datasets and tasks for drug discovery and development. *NeurIPS Datasets and Benchmarks Track*, 2021.

[29] A. K. Debnath, R. L. Lopez de Compadre, G. Debnath, A. J. Shusterman, and C. Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, 34(2):786–797, 1991.

[30] C. Morris, N. M. Kriege, F. Bause, K. Kersting, P. Mutzel, and M. Neumann. TUDataset: a collection of benchmark datasets for learning with graphs. *ICML Workshop on Graph Representation Learning and Beyond*, 2020.

[31] G. W. Bemis and M. A. Murcko. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996.

[32] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning* (*ICML*), PMLR 70:1321–1330, 2017.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- Additionalfile3.csv
- Additionalfile4.csv
- Additionalfile2.csv