Discover, Ground, Explain: Designing Intrinsically Interpretable GNNs

Anonymous authors
Paper under double-blind review

Abstract

Molecular graph models can achieve high predictive performance while their attributions poorly localize the substructures underlying the target property. We introduce Graph Variational Motif Prior (GraphVMP), a variational GNN for intrinsic graph explainability that aligns a dedicated latent partition with graphlet descriptors through maximum mean discrepancy (MMD). To learn structural references without concept supervision, we design a novel unsupervised grouping model requiring only upper bounds on subgraph size and the number of distinct motif types as prior information. We also introduce MotifCausalShift, a controlled structural out-of-distribution (OOD) benchmark that varies spurious motif correlations and introduces unseen test-time distractors. On this benchmark, grouping recovers all four planted training motifs and isolates an unseen Star motif without a corresponding learned template. Without annotations specifying which motifs are causal, GraphVMP achieves $97.0 \pm 1.0\%$ accuracy and $85.0 \pm 6.0\%$ explanation ROC-AUC in the primary MotifCausalShift stress test, improving explanation ROC-AUC by 29.6 percentage points over the baseline without MMD. Following causal-motif replacement, node ROC-AUC remains $85.7 \pm 13.7\%$, compared to $56.0 \pm 19.4\%$ without alignment. These results support structural discovery and explicit latent alignment as an approach to designing intrinsic graph explanations whose localization remains effective under controlled structural shifts.

1 Introduction

In molecular property prediction, a model may exploit a substructure that correlates with the label without reflecting the mechanism of interest McCloskey et al. (2019). This motivates a central question in the design of intrinsic explainability: how can we ground a graph predictor's explanatory representations in structural concepts and assess whether its decisions depend on those concepts? We propose GraphVMP, an architecture that grounds latent representations in structural descriptors through maximum mean discrepancy (MMD) alignment. We further introduce an unsupervised grouping approach to discover a shared vocabulary of subgraphs from which these descriptors can be constructed.

Post-hoc explanation methods Ying et al. (2019); Yuan et al. (2021); Luo et al. (2020) analyse a trained predictor but do not directly constrain how its explanatory representations are learned. Intrinsically interpretable approaches incorporate explanatory mechanisms through information bottlenecks Wu et al. (2020); Miao et al. (2022) and prototype-based reasoning Zhang et al. (2022); Seo et al. (2023). DIR Wu et al. (2022) constructs interventional distributions to encourage prediction-stable rationales, while TopInG Xin et al. (2025) introduces topological constraints to distinguish rationale subgraphs from irrelevant structure. These approaches impose different inductive biases on the relationship between graph structure and prediction. When task-relevant and spurious substructures are correlated, however, predictive accuracy alone does not establish which structure drives the decision or whether the explanation reflects that dependence. Our architectural approach addresses this question by explicitly aligning explanatory latent components with structural descriptors and evaluating the model's response to controlled changes in the corresponding subgraphs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

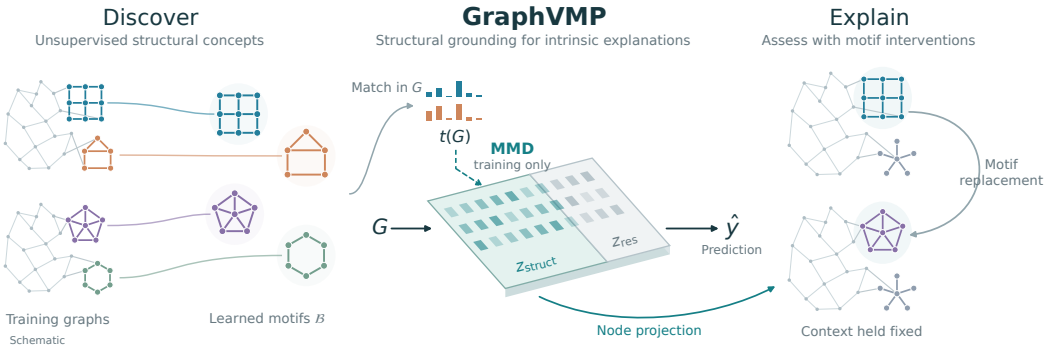

Figure 1: Unsupervised grouping discovers motif templates that supply graph-specific graphlet targets for MMD grounding. Controlled motif replacement probes whether node-level explanations follow the new motif.

GraphVMP (Variational Motif Prior) partitions the latent representation into structurally grounded and unconstrained residual components. Graphlet histograms provide compact descriptions of reference motif topology, and an MMD objective encourages the designated structural channels to align with these descriptors.

The grouping stage learns graph-specific node partitions and a shared vocabulary of subgraph templates, guided by bounds on subgraph size and vocabulary capacity. It requires neither motif identities nor node-level annotations. Graphlet descriptors computed from the learned templates provide structural targets for GraphVMP.

We introduce MotifCausalShift to evaluate robustness and explanatory localization under controlled structural shifts. In its weighted spurious setting, GraphVMP achieves $97.0 \pm 1.0\%$ accuracy and $85.0 \pm 6.0\%$ explanation ROC-AUC, compared with $55.4 \pm 17.2\%$ ROC-AUC for the baseline without MMD. Additional experiments evaluate predictive performance and explanation quality on standard in-distribution benchmarks.

Our main contributions are summarized as follows:

- **Unsupervised structural discovery.** We introduce a novel grouping model that learns shared subgraph templates and graph-specific node partitions without class labels or motif annotations. The learned templates supply graphlet descriptors.
- **Structural grounding for intrinsic explainability.** We introduce GraphVMP, a variational GNN whose explanations are represented by dedicated latent components. MMD regularisation grounds these components using graphlet descriptors computed from subgraph templates discovered through unsupervised grouping.
- **MotifCausalShift.** We introduce a synthetic graph benchmark with controlled spurious motif–label correlations and unseen test-time distractors. Targeted motif replacement and deletion enable evaluation of how predictions and explanations respond to changes in causal and spurious structure.

2 Related Work

Post-hoc and intrinsically interpretable GNN explanations. Post-hoc methods identify important nodes, edges, or subgraphs in trained GNNs Ying et al. (2019); Yuan et al. (2021); Luo et al. (2020); Vu & Thai (2020); Schlichtkrull et al. (2021), without constraining how their representations are learned. Intrinsic approaches integrate explanation mechanisms into the predictive model through information bottleneck objectives Yu et al. (2021), prototype-based reasoning Zhang et al. (2022); Seo et al. (2023), or symbolic rule induction Ragno et al. (2026). However, incorporating an explanation mechanism does not guarantee faithfulness itself. Azzolin et al. (2026) show that self-explainable GNNs can achieve optimal predictive performance while producing explanations unrelated to their inference

process, and Tai et al. (2025) further show that self-interpretable GNNs can overemphasize task-irrelevant edges despite high predictive accuracy, resulting in inconsistent and inaccurate explanations. These findings highlight that predictive accuracy alone does not ensure explanation faithfulness or accurate localization of task-relevant structures. GraphVMP introduces explicit structural grounding by aligning a dedicated latent partition with graphlet descriptors through MMD, using subgraph templates discovered by unsupervised grouping as structural references.

Rationales and graph OOD generalization. Graph out-of-distribution (OOD) methods seek predictive structures that remain stable under distribution shifts. GSAT Miao et al. (2022) uses an information bottleneck with stochastic attention to suppress task-irrelevant structure; DIR Wu et al. (2022) constructs interventional distributions to learn invariant rationales; and CIGA Chen et al. (2022) extracts subgraphs that preserve invariant intra-class information. PrunE Yao et al. (2026) suppresses spurious edges through graph-size constraints and probability alignment, while InfoIGL Mao et al. (2026) combines information compression with multi-level contrastive learning to retain invariant features. These approaches impose different assumptions and objectives for separating predictive structure from spurious correlations. When causal and spurious motifs are indistinguishable from training observations, predictive fit alone cannot determine which is causal. GraphVMP provides structural grounding by discovering a shared motif vocabulary without class labels or motif annotations and aligning explanatory latent components with its descriptors.

Subgraph discovery and dictionary learning. Online Graph Dictionary Learning Vincent-Cuaz et al. (2021) approximates each graph with a single mixture of learned structural atoms, without explicitly decomposing it into separate motif instances. In contrast, our model learns this decomposition: it identifies node groups and matches each group separately to a shared vocabulary of subgraph templates. MGSSL Zhang et al. (2021) obtains local motifs through chemistry-specific fragmentation; our groups and template structures are learned from graph topology without predefined fragmentation rules or motif annotations.

3 Methodology

We design GraphVMP for intrinsic graph explainability by grounding a dedicated variational latent partition in structural concepts learned from data. Our unsupervised grouping model discovers subgraph templates, whose graphlet histograms Shervashidze et al. (2009) provide the targets for MMD alignment Gretton et al. (2012). Training proceeds in two stages: we first train the unsupervised grouping model to obtain the shared motif vocabulary, then train the variational graph model using structural targets derived from this vocabulary.

3.1 Problem setup and explanation target

Let $\mathcal{D} = \{(G_n, y_n)\}_{n=1}^{N}$ be a graph classification dataset, where $G = (V, E, X)$ has node features $X \in \mathbb{R}^{|V| \times d}$ and label $y \in \{1, \ldots, C\}$. Our goal is to learn a graph classifier whose explanatory latent components align with structural references discovered without class labels or motif annotations.

Unsupervised grouping discovers a shared vocabulary $\mathcal{B} = \{B_m\}_{m=1}^{M}$ of template adjacency matrices. Their graphlet histograms define the structural references $\mathcal{S} = \{h(B_m)\}_{m=1}^{M}$, where h is the graphlet descriptor map. An explanation comprises the structural latent components and their supporting input subgraphs.

Non-identifiability from predictive fit. When causal and spurious motifs perfectly co-occur during training, predictors relying on either can produce identical predictions. Predictive accuracy alone therefore cannot establish which substructures drive a model's decisions. GraphVMP grounds explanatory latent components in discovered subgraph concepts, while controlled motif interventions assess whether the resulting explanations correspond to the model's predictive behavior.

3.2 Graph encoder and latent tokenization

We first map the variable-size input graph to a fixed-size latent representation. A graph encoder produces node embeddings

$$H = \mathrm{Enc}_\phi(G) \in \mathbb{R}^{|V| \times d_h}.$$

In our experiments, the encoder is two blocks of GCN. To obtain a fixed number of latent tokens, we subsequently apply a transformer-based pooling in our implementation

$$U = \mathrm{Pool}(H) \in \mathbb{R}^{\kappa \times d_h},$$

where κ is the maximum number of retained graph tokens. The pooling operation also induces a node-to-token support matrix

$$P(G) \in \mathbb{R}^{|V| \times \kappa},$$

so that each pooled token can be related back to the original graph. Each deterministic token u_i parameterizes a Gaussian latent token:

$$\mu_i = g_\mu(u_i), \qquad \log \sigma_i^2 = g_\sigma(u_i),$$

and

$$q_\phi(z_i \mid U) = \mathcal{N}\big(z_i; \mu_i, \mathrm{diag}(\sigma_i^2)\big).$$

The full variational approximation is assumed to be a mean-field model

$$q_\phi(z \mid G) = q_\phi(z \mid U(G)) = \prod_{i=1}^{\kappa} q_\phi(z_i \mid U). \tag{1}$$

Motif-structured latent prior We use a learnable K-component Gaussian mixture prior over latent tokens:

$$p_\theta(z_i) = \sum_{m=1}^{K} \pi_m \mathcal{N}(z_i; \mu_m, \Sigma_m), \qquad p_\theta(z) = \prod_{i=1}^{\kappa} p_\theta(z_i). \tag{2}$$

Here, π_m, μ_m, and Σ_m are the mixture weights, means, and covariances, with $\pi_m \geq 0$ and $\sum_{m=1}^{K} \pi_m = 1$.

Structural grounding and prediction For each graph G, a detector identifies the templates present from the dictionary $\mathcal{B} = \{B_m\}_{m=1}^{M}$ learned through unsupervised grouping (subsection 3.3). Their graphlet descriptors are assembled into the graph-specific structural target $t(G)$. We align structural latent vectors $\{z_i^s\}_{i=1}^{n_z}$ with reference vectors $\{t_j\}_{j=1}^{n_t}$ constructed from these targets using the biased squared MMD estimator:

$$\begin{aligned} \mathcal{L}_{\mathrm{MMD}} = \frac{1}{n_z^2} \sum_{i,i'=1}^{n_z} k(z_i^s, z_{i'}^s) + \frac{1}{n_t^2} \sum_{j,j'=1}^{n_t} k(t_j, t_{j'}) \\ - \frac{2}{n_z n_t} \sum_{i=1}^{n_z} \sum_{j=1}^{n_t} k(z_i^s, t_j). \end{aligned} \tag{3}$$

where k is a positive-definite kernel, $z_i^s, t_j \in \mathbb{R}^r$, and n_z, n_t denote the numbers of structural latent and reference vectors. The classifier predicts from the full latent representation, $\hat{y} = f_\theta(z)$.

3.3 Learning a dictionary of subgraphs

We learn a shared dictionary of subgraph templates together with graph-specific node groups. Gromov–Wasserstein matching aligns each group with the templates, while template-based reconstruction provides feedback for iterative group refinement. For an n-node graph with adjacency A, we allow up to K_g groups and use M templates; template m has q_m nodes, bounded by $k_{\max}$.

Grouping and matching. A three-layer GIN (Xu et al., 2019) encodes constant node features and Laplacian positional encodings (Dwivedi et al., 2023) into H. An MLP initializes logits L and memberships $S = \mathrm{softmax}_{\mathrm{row}}(L) \in [0,1]^{n \times K_g}$. Template logits $C_m \in \mathbb{R}^{q_m \times q_m}$ parameterize symmetric, zero-diagonal adjacencies $B_m = \sigma((C_m + C_m^\top)/2) \odot (\mathbf{1}_{q_m}\mathbf{1}_{q_m}^\top - I_{q_m})$, where σ is the sigmoid. Current memberships define groups $V_k = \{i : \arg\max_j S_{ij} = k\}$. For each eligible group of size $n_k = |V_k| \in [3, k_{\max}]$, unrolled entropic Gromov–Wasserstein (GW) matching (Peyré et al., 2016) compares $A[V_k, V_k]$ with every B_m using uniform node masses, yielding couplings Π_{km} and squared-loss distortions d_{km}. Embedding $n_k \Pi_{km}$ into $P_{km} \in \mathbb{R}^{n \times q_m}$ with zero rows outside V_k, we obtain

$$\begin{aligned} \alpha_{km} &= \frac{\exp(-d_{km}/\tau_a)}{\sum_{j=1}^{M} \exp(-d_{kj}/\tau_a)}, \\ Q_k &= \sum_{m=1}^{M} \alpha_{km} P_{km} B_m P_{km}^\top. \end{aligned} \tag{4}$$

where $\tau_{\mathrm{a}} > 0$ is a temperature; ineligible groups have $Q_k = 0$.

Reconstruction and learning. The decoder combines group contributions with a background model:

$$\begin{aligned} \widehat{A}_{ij} &= 1 - (1 - p_{ij}^{\mathrm{bg}}) \prod_{k=1}^{K_g} (1 - S_{ik} S_{jk} Q_{k,ij}), \\ p_{ij}^{\mathrm{bg}} &= \sigma(b + u_i + u_j). \end{aligned} \tag{5}$$

where u_i is a learned affine function of H_i. At each of R refinement rounds, an MLP updates

$$L \leftarrow L + f_{\mathrm{ref}}([H, S, e]), \tag{6}$$

where $e_{ik} = \frac{\sum_j S_{jk} |A_{ij} - Q_{k,ij}|}{\sum_j S_{jk} + \epsilon}$ and $\epsilon > 0$ ensures numerical stability.

Memberships and matching are recomputed after each update. The unsupervised objective is

$$\mathcal{L}_{\mathrm{disc}} = \lambda_{\mathrm{rec}} \, \mathrm{BCE}_{i \neq j}(A, \widehat{A}) + \lambda_{\mathrm{fit}} \mathcal{L}_{\mathrm{fit}} + \mathcal{L}_{\mathrm{group}}, \tag{7}$$

Here, $\mathcal{L}_{\mathrm{fit}}$ averages $\min_m d_{km}$ over eligible groups. The grouping term $\mathcal{L}_{\mathrm{group}}$ combines weighted normalized-cut and orthogonality penalties (Bianchi et al., 2020) with a squared penalty on group mass exceeding $k_{\max}$. Warm-up uses only $\mathcal{L}_{\mathrm{group}}$. Templates are then initialized from training groups: the first uniformly, and subsequent ones with probability proportional to squared GW dissimilarity to the nearest selected template. At epoch boundaries, training structures scored by WL-hash frequency times edge count can replace lower-scoring templates, allowing size changes. Gradients propagate through transport, soft template weights, and reconstruction; hard group extraction and template replacement are discrete.

3.4 Training objective

The GraphVMP predictor is trained with classification, latent regularization, and MMD-based structural grounding:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \beta \, \mathcal{L}_{\mathrm{KL}} + \gamma \, \mathcal{L}_{\mathrm{MMD}}, \tag{8}$$

where $\mathcal{L}_{\mathrm{MMD}}$ is defined in Equation 3. For a graph-label pair (G, y), the classification loss is

$$\mathcal{L}_{\mathrm{cls}} = \mathbb{E}_{q_\phi(z|G)}[-\log p_\theta(y \mid z)].$$

The KL term regularizes the posterior toward the motif-structured prior:

$$\mathcal{L}_{\mathrm{KL}} = \mathbb{E}_{q_\phi(z|G)} \left[\log q_\phi(z \mid G) - \log p_\theta(z)\right].$$

The weights β and γ control latent regularization and structural grounding, respectively.

Algorithm 1 Subgraph discovery and target construction

Require: Unlabelled training graphs $\{G_n\}_{n=1}^{N}$
Warm up with $\mathcal{L}_{\text{group}}$
Initialize $\mathcal{B} = \{B_m\}_{m=1}^{M}$ by GW diversity
for each training epoch **do**
  for each graph G **do**
    Encode G; initialize L, S
    Run R matching/refinement rounds
    Rematch; decode $\widehat{A}$ with background
    Update model using $\mathcal{L}_{\text{disc}}$
  end for
  Refresh templates via WL proposals
end for
Compute template descriptors $\{h(B_m)\}_{m=1}^{M}$
for each training graph G_n **do**
  Detect templates present in G_n using $\mathcal{B}$
  Construct $t(G_n)$ from their descriptors
end for
return grouping model, $\mathcal{B}$, $\{t(G_n)\}_{n=1}^{N}$

Algorithm 2 GraphVMP training

Require: Training data $\mathcal{D}$; fixed weight γ
Require: Targets $\{t(G_n)\}_{n=1}^{N}$ from Algorithm 1
Initialize θ, ϕ; set $\beta \leftarrow 10^{-3}$
for each training batch $\mathcal{D}_b$ **do**
  for each $(G, y) \in \mathcal{D}_b$ **do**
    $z(G) \sim q_\phi(z \mid G)$
    $\hat{y}(G) = f_\theta(z(G))$
  end for
  Compute batch losses $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{KL}}$
  Compute $\mathcal{L}_{\text{MMD}}$ from structural latent vectors and batch targets using Eq. 3
  $\mathcal{L} = \mathcal{L}_{\text{cls}} + \beta\mathcal{L}_{\text{KL}} + \gamma\mathcal{L}_{\text{MMD}}$
  Update θ, ϕ using $\nabla\mathcal{L}$
end for
return trained GraphVMP model

4 Experimental Results

We evaluate whether structural discovery and MMD alignment can support intrinsic graph explainability under distribution shift while preserving predictive accuracy. We first assess unsupervised subgraph discovery: our novel grouping model recovers the training motifs and isolates the unseen Star motif at test time, despite having no learned Star template. We then evaluate structural grounding through OOD stress-tests and controlled motif interventions. Following causal-motif replacement, MMD-aligned explanations retain $85.7 \pm 13.7\%$ node ROC-AUC, compared to $56.0 \pm 19.4\%$ without alignment, showing that the localization advantage persists when the causal motif changes. Finally, standard graph benchmarks assess predictive performance and node-level explanation quality.

4.1 Datasets and Implementation Details

Datasets. We evaluate GraphVMP on four benchmark datasets: BA2Motifs Yuan et al. (2020), BAMultiShapes Azzolin et al. (2023), Mutagenicity Morris et al. (2020), and Benzene Sanchez-Lengeling et al. (2020). BA2Motifs and BAMultiShapes are synthetic datasets with planted explanatory motifs; Mutagenicity and Benzene are molecular classification datasets with substructure annotations for explanation evaluation. We additionally introduce *MotifCausalShift*, a controlled synthetic benchmark testing robustness and causal-subgraph localization under changes in spurious motif–label associations and unseen structural distractors. Standard dataset descriptions appear in Appendix B; *MotifCausalShift* is detailed in Section 4.3.

Implementation Details. We use the Adam optimizer with a learning rate of 10^{-3} and keep the KL regularization weight fixed at $\beta = 10^{-3}$ throughout training. We vary the MMD regularization weight $\gamma \in \{0.25, 0.5, 0.75, 1.0\}$ on MotifCausalShift, while setting $\gamma = 1.0$ for all other datasets. All experiments are conducted on an NVIDIA L40S GPU.

4.2 Can the structural descriptor be discovered rather than specified?

On MotifCausalShift, we evaluate whether unsupervised grouping recovers the planted subgraph templates and their graphlet descriptors without class labels or motif annotations.

The grouping model uses only adjacency and Laplacian positional encodings. Ground-truth motif information is used solely for evaluation (Table 1). For each motif instance, we select the group with maximum node overlap and measure *recall* (the fraction of motif nodes recovered)

Table 1: Unsupervised structure discovery on MotifCausalShift, three seeds. Left: every seed recovers all four planted motifs exactly, with graphlet vectors identical to the hand-specified descriptor. Right: localization of motif instances on the held-out split, measured from the node partition alone. Star appears only at test time and therefore has no learned template, yet is still isolated.

(a) Discovered templates (training split).

Motif	Nodes	Edges	Graphlet vector
House	5	6	$[6, 9, 1, 10, 2, 1]$
Cycle	6	6	$[6, 6, 0, 6, 0, 0]$
Wheel	6	10	$[10, 25, 5, 45, 15, 5]$
Grid	9	12	$[12, 22, 0, 40, 8, 4]$

(b) Localization, 500 held-out graphs.

Motif	Recall	Precision
Grid	0.80 ± 0.04	$\mathbf{0.98} \pm 0.01$
Wheel	0.98 ± 0.01	0.95 ± 0.01
Star†	$\mathbf{0.99} \pm 0.00$	0.88 ± 0.01

Table 2: OOD test accuracy (%) on MotifCausalShift (Exp. 1) and Balanced Spurious (Exp. 2).

Method	**Exp. 1**	**Exp. 2**
Baseline	60.0	81.0
GraphVMP	**80.0**	**94.0**

Table 3: Weighted Spurious Stress-Test (Exp. 3): test accuracy and explanation ROC-AUC (%).

Method	**Acc. ↑**	**ROC-AUC ↑**
Baseline	90.0 ± 10.0	55.4 ± 17.2
DIR	63.2 ± 1.7	31.5 ± 14.2
GraphVMP	$\mathbf{97.0 \pm 1.0}$	$\mathbf{85.0 \pm 6.0}$

and *precision* (the fraction of selected-group nodes belonging to the motif). Planted motifs are recovered by enumerating induced copies of each learned template; we report the resulting candidate sets.

On held-out graphs with unseen motif combinations, the model isolates the test-only Star motif with recall 0.99 ± 0.00, despite having no corresponding learned template. Wheel and Star groups absorb adjacent background nodes, reducing precision, whereas Grid is recovered with high precision but incomplete coverage.

4.3 Structural OOD Evaluation

We evaluate GraphVMP under three controlled structural OOD settings. Experiment 1 reverses spurious associations between familiar motifs and class labels. Experiments 2 and 3 replace training distractors with an unseen Star, following balanced or class-correlated distractor sampling during training, respectively. Experiment 3 is the primary stress test: MMD alignment improves both classification accuracy and node ROC-AUC over the baseline without MMD, whose high predictive accuracy coexists with poor localization of the planted causal motifs.

Experiment 1: MotifCausalShift. Each graph has a fixed total node count and contains two motifs: grid + house (Class 0) or wheel + cycle (Class 1) during training. At test time, these become grid + cycle and wheel + house, respectively. This reverses the spurious motif–label associations while preserving the grid/wheel label relationship. GraphVMP achieves 80.0% test accuracy versus 60.0% for the baseline without MMD (Table 2).

Experiment 2: Unseen-Distractor Robustness (Balanced Training). In Experiment 1, causal and spurious motifs are equally predictive during training, so training accuracy cannot distinguish causal reliance from shortcuts. Experiment 2 removes this spurious label signal by balancing house and cycle (50/50) within each class. MMD alignment covers both causal and spurious motifs. At test time, spurious motifs are replaced by an unseen Star while causal motifs remain unchanged, testing robustness to novel spurious structure. GraphVMP achieves 94.0% test accuracy versus 81.0% for the baseline without MMD (Table 2), indicating improved robustness to the replacement of balanced training distractors with an unseen Star motif.

Table 4: Counterfactual motif intervention on the 500 OOD test graphs (8 seeds).

	γ=1 (with structural prior)			γ=0 (no prior)		
intervention	changed (%)	p_{keep}	node AUC	changed (%)	p_{keep}	node AUC
none (reference)	—	1.00	$\mathbf{85.9 \pm 13.8}$	—	1.00	55.8 ± 19.6
swap causal (Grid $\leftrightarrow$ Wheel)	81.7 ± 17.8	0.18	$\mathbf{85.7 \pm 13.7}$	83.6 ± 10.5	0.17	56.0 ± 19.4
delete nuisance	6.4 ± 5.1	0.94	$\mathbf{86.2 \pm 13.4}$	5.4 ± 3.4	0.95	54.1 ± 18.9
delete causal	42.3 ± 7.3	0.58	—	42.4 ± 3.7	0.58	—

Experiment 3: Unseen-Distractor Robustness (Spurious Training Correlations) (Primary). As in Experiment 2, MMD targets cover all four motifs without specifying which are causal. House now occurs in 80% of Class 0 samples and Cycle in 80% of Class 1 samples, testing causal-subgraph recovery under strong shortcut pressure without privileged causal-motif annotations. At test time, spurious motifs are replaced by an unseen Star while causal motifs remain unchanged.

The baseline achieves $90.0 \pm 10.0\%$ accuracy but near-chance explanation ROC-AUC ($55.4 \pm 17.2\%$), indicating poor causal-subgraph localization. DIR performs worse, with $63.2 \pm 1.7\%$ accuracy and $31.5 \pm 14.2\%$ ROC-AUC. GraphVMP ($\gamma = 1.0$) achieves $97.0 \pm 1.0\%$ accuracy and $85.0 \pm 6.0\%$ ROC-AUC, improving explanation ROC-AUC by 29.6 percentage points over the baseline. Full per-seed results appear in Appendix C.

Comparison with DIR. We evaluate DIR Wu et al. (2022) in Experiment 3, where class-correlated spurious motifs are replaced by an unseen Star at test time. DIR achieves lower test accuracy and explanation ROC-AUC than the unconstrained baseline (Table 3), indicating limited robustness under this structural shift.

4.4 Motif Intervention Analysis

We evaluate prediction sensitivity and node-level localization under motif replacement and deletion using prediction changes and node ROC-AUC. Both variants in Table 4 show similar aggregate prediction-flip rates, with greater sensitivity to causal swaps than to nuisance deletion. MMD increases reference node ROC-AUC from 55.8 to 85.9. After replacing one familiar causal motif with another while keeping the unseen Star distractor fixed, node ROC-AUC remains 85.7 with MMD, compared with 56.0 without it.

4.5 Predictive Performance on Standard Benchmarks

To assess whether GraphVMP remains competitive in predictive accuracy, we compare it with intrinsically interpretable GNN baselines on standard benchmarks (Table 5).

GraphVMP reaches 100.0% accuracy on BA2Motifs and 99.0% on BAMultiShapes. On Mutagenicity, GraphVMP achieves 80.0% accuracy; this relatively lower performance is expected, as Mutagenicity does not provide clean ground-truth substructures suitable for weak structural supervision, placing the method at a disadvantage. For the Benzene dataset, which contains a single clear causal motif, most reasonable GNNs can already achieve near-perfect performance. As such, Benzene serves primarily as a sanity check rather than a benchmark for demonstrating improvements. Results suggest that structural alignment preserve accuracy while improving interpretability. The benefits of MMD-based structural alignment become most evident in the presence of spurious correlations, as illustrated by the MotifCausalShift setting.

4.6 Node-Level Localization

GraphVMP's explanatory representations are structurally grounded latent components. For comparison with node-level baselines, we project these components to node importance

Table 5: Classification accuracy (%).

Method	BA2Motifs	BAMulti Shapes	Mutagenicity	Benzene
LogiX-GIN	**100.00 ± 0.00**	**100.00 ± 0.00**	79.31 ± 1.25	N/A
PiGNN	99.89 ± 0.31	85.40 ± 5.08	**82.39 ± 1.68**	N/A
GIB	**100.00 ± 0.00**	97.60 ± 2.06	80.14 ± 0.98	N/A
KerGNN	98.80 ± 0.75	83.20 ± 1.94	73.32 ± 2.93	N/A
TopInG	**100.00 ± 0.00**	N/A	N/A	**100.00 ± 0.00**
GraphVMP	**100.00 ± 0.00**	99.00 ± 1.00	80.00 ± 1.00	90.00 ± 1.00

Table 6: Node ROC-AUC (%).

Method	BA2Motifs	Mutagenicity	Benzene
GSAT	98.85 ± 0.47	99.38 ± 0.25	91.57 ± 1.48
DIR	82.78 ± 10.97	64.44 ± 28.81	54.08 ± 13.75
CIGA	N/A	N/A	N/A
TopInG	**99.57 ± 0.60**	95.79 ± 1.93	**98.22 ± 0.92**
GraphVMP	95.8 ± 3.6	**100.0 ± 0.0**	91.9 ± 2.3

Values are mean ± standard deviation over seeds. Bold: highest mean per column; N/A: unavailable. GraphVMP node scores follow Eq. 9.

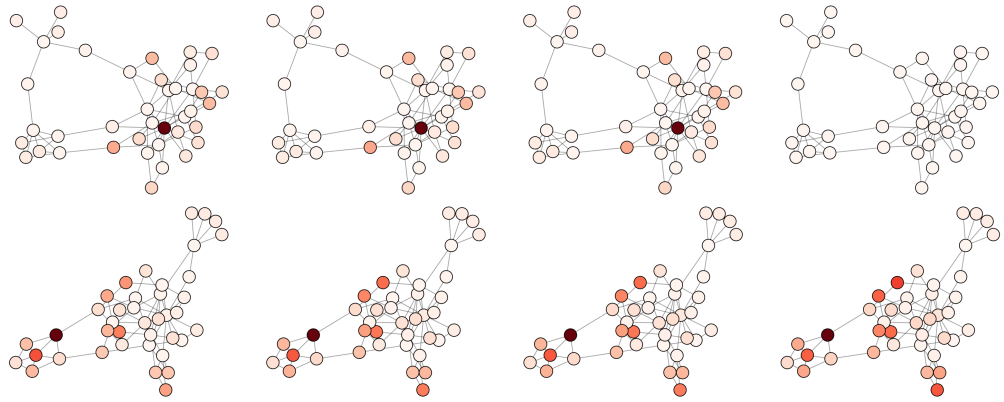

Figure 2: Node-level localization of structural latent channels on MotifCausalShift. Rows show examples without MMD (top) and with MMD (bottom). Each column visualizes one channel of S (Eq. (9)), obtained by weighting Transformer node assignments with structural latent activations. Darker red indicates higher importance. With MMD (bottom), the highest node-importance scores align with the wheel motif.

scores S:

$$S = \mathrm{ReLU}(Z_{\mathrm{struct}})^{\top} A, \tag{9}$$

where Z_{struct} contains the structural latent activations and A contains the Transformer node-assignment weights.

Table 6 reports ROC-AUC of these node importance scores against ground-truth node labels. GraphVMP achieves $100.0 \pm 0.0\%$ on Mutagenicity — the only method to do so — and competitive results on BA2Motifs ($95.8 \pm 3.6\%$) and Benzene ($91.9 \pm 2.3\%$). The gap behind GSAT and TopInG on BA2Motifs reflects the coarseness of the projection in Equation 9: methods that optimize node-level masks end-to-end have a structural advantage on this metric, as the MotifCausalShift results demonstrate. Qualitative results are shown in Figure 2.

5 Conclusion

We introduced GraphVMP, which grounds explanatory latent components through MMD alignment, together with a novel unsupervised grouping method for discovering structural concepts. The grouping model recovers training motifs and isolates unseen Star subgraphs without motif annotations. Under structural OOD shifts, MMD improves causal-subgraph localisation, and its ROC-AUC advantage persists after causal-motif replacement. Similar aggregate prediction responses with and without MMD highlight that predictive sensitivity alone does not establish explanation quality. Together, these findings support designing intrinsic graph explainability around learned structural concepts and assessing their grounding through controlled interventions.

AI use statement

We used generative AI tools to assist with literature search and editing the manuscript. The authors take responsibility for the final manuscript, references, methods, and reported results.

References

Steve Azzolin, Antonio Longa, Pietro Barbiero, Pietro Lio, and Andrea Passerini. Global Explainability of GNNs via Logic Combination of Learned Concepts. In *ICLR*, 2023.

Steve Azzolin, Stefano Teso, Bruno Lepri, Andrea Passerini, and Sagar Malhotra. GNN Explanations That Do Not Explain and How to Find Them. In *ICLR*, 2026.

Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral Clustering with Graph Neural Networks for Graph Pooling. In *ICML*, 2020.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, Kaili Ma, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning Causally Invariant Representations for Out-of-Distribution Generalization on Graphs. In *NeurIPS*, 2022.

Asim Kumar Debnath, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman, and Corwin Hansch. Structure-activity Relationship of Mutagenic Aromatic and Heteroaromatic Nitro Compounds. Correlation with Molecular Orbital Energies and Hydrophobicity. *J. Med. Chem.*, 1991.

Vijay Prakash Dwivedi, Chaitanya K Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking Graph Neural Networks. *JMLR*, 2023.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *JMLR*, 2012.

Jeroen Kazius, Ross McGuire, and Roberta Bursi. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.*, 2005.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized Explainer for Graph Neural Network. *NeurIPS*, 2020.

Wenyu Mao, Jiancan Wu, Haoyang Liu, Yongduo Sui, and Xiang Wang. Invariant Graph Learning Meets Information Bottleneck for Out-of-Distribution Generalization. *Front. Comput. Sci.*, 2026.

Kevin McCloskey, Ankur Taly, Federico Monti, Michael P Brenner, and Lucy J Colwell. Using Attribution to Decode Binding Mechanism in Neural Network Models for Chemistry. *PNAS*, 2019.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism. In *ICML*, 2022.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A Collection of Benchmark Datasets for Learning with Graphs. In *ICMLW)*, 2020.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *ICML*, 2016.

Alessio Ragno, Marc Plantevit, and Céline Robardet. On Logic-based Self-Explainable Graph Neural Networks. *NeurIPS*, 2026.

Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Wang, Wesley Qian, Kevin McCloskey, Lucy Colwell, and Alexander Wiltschko. Evaluating Attribution for Graph Neural Networks. *NeurIPS*, 2020.

Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking. In *ICLR*, 2021.

Sangwoo Seo, Sungwon Kim, and Chanyoung Park. Interpretable Prototype-based Graph Information Bottleneck. *NeurIPS*, 2023.

Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient Graphlet Kernels for Large Graph Comparison. In *AISTATS*, 2009.

Teague Sterling and John J Irwin. ZINC 15–ligand Discovery for Everyone. *J. Chem. Inf. Model.*, 2015.

Wenxin Tai, Ting Zhong, Goce Trajcevski, and Fan Zhou. Redundancy Undermines the Trustworthiness of Self-interpretable GNNs. In *ICML*, 2025.

Cédric Vincent-Cuaz, Titouan Vayer, Rémi Flamary, Marco Corneli, and Nicolas Courty. Online Graph Dictionary Learning. In *ICML*, 2021.

Minh Vu and My T Thai. Pgm-explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks. *NeurIPS*, 2020.

Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. Graph Information Bottleneck. *NeurIPS*, 2020.

Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering Invariant Rationales for Graph Neural Networks. In *ICLR*, 2022.

Cheng Xin, Fan Xu, Xin Ding, Jie Gao, and Jiaxin Ding. TopInG: Topologically Interpretable Graph Learning via Persistent Rationale Filtration. In *ICML*, 2025.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful Are Graph Neural Networks? In *ICLR*, 2019.

Tianjun Yao, Haoxuan Li, Yongqiang Chen, Tongliang Liu, Le Song, Eric Xing, and Zhiqiang Shen. Pruning Spurious Subgraphs for Graph Out-of-Distribution Generalization. *NeurIPS*, 2026.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating Explanations for Graph Neural Networks. In *NeurIPS*, 2019.

Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. Graph Information Bottleneck for Subgraph Recognition. In *ICLR*, 2021.

Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. XGNN: Towards Model-level Explanations of Graph Neural Networks. In *ACM SIGKDD*, 2020.

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On Explainability of Graph Neural Networks via Subgraph Explorations. In *ICML*, 2021.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based Graph Self-Supervised Learning for Molecular Property Prediction. *NeurIPS*, 2021.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. Protgnn: Towards Self-explaining Graph Neural Networks. In *AAAI*, 2022.

A Datasets

BA2Motifs is an extension of some popular synthetic benchmarks Yuan et al. (2020) aimed to assess the ability of a Global Explainer. In particular, the dataset is composed of 1,000 Barabasi-Albert (BA) graphs with attached in random positions the following network motifs: house, grid, wheel. Class 0 contains plain BA graphs and BA graphs enriched with a house, a grid, a wheel, or the three motifs together. Class 1 contains BA graphs enriched with a house and a grid, a house and a wheel, or a wheel and a grid.

Table 7: Statistics of the public benchmark datasets. Edges are counted once per undirected pair, excluding self-loops. Averages are rounded to one decimal place.

Dataset	MUTAG	Mutagenicity	BA2Motifs	BAMultiShapes	Benzene
Graphs	188	4,337	1,000	1,000	12,000
Classes	2	2	2	2	2
Max nodes	28	417	25	40	25
Avg nodes	17.9	30.3	25.0	40.0	20.6
Avg edges	19.8	30.8	25.5	43.7	21.8
GT type*	Chemical	Chemical	Planted	Planted	Ring-based

*Chemical: protocol-dependent molecular substructure annotations. Planted: motifs specified by the synthetic generator. Ring-based: benzene-ring membership. Dataset sources: MUTAG Debnath et al. (1991); Mutagenicity Kazius et al. (2005); BA2Motifs Luo et al. (2020); BAMultiShapes Azzolin et al. (2023); Benzene Sanchez-Lengeling et al. (2020). TU-release statistics follow Morris et al. (2020).

BAMultiShapes Azzolin et al. (2023) is a synthetic benchmark with logical combinations of graph concepts. It consists of 1,000 Barabasi-Albert (BA) graphs augmented with house, grid, and wheel motifs attached at random positions. Class 0 contains plain BA graphs, single motifs, or all three motifs together, while Class 1 contains pairwise combinations of motifs.

Benzene Dataset Sanchez-Lengeling et al. (2020) is a binary molecular graph classification dataset containing 12,000 compounds sampled from the ZINC15 database Sterling & Irwin (2015). The task is to predict whether a molecule contains at least one benzene ring. Molecules are represented as graphs with atoms as nodes and chemical bonds as edges. In explainability evaluations, the atoms forming the benzene ring are treated as ground-truth explanatory substructures, and multiple disconnected benzene rings are considered as separate explanations.

MUTAG a molecular graph classification dataset of 188 nitroaromatic compounds. Each graph is a molecule, with atoms as nodes and bonds as edges, and the task is to predict whether the compound is mutagenic to Salmonella typhimurium. In explainability papers, MUTAG is often treated as having ground-truth explanatory substructures for the mutagenic class, typically certain nitro-containing aromatic functional groups. Those substructures are used as the "relevant rationale" when evaluating explanations Debnath et al. (1991).

B Details of our architecture

Transformer-based pooling: The original sort pooling layer was removed and replaced with a Transformer pooling mechanism. Unlike sort pooling, which deterministically ranks nodes based on feature magnitudes, Transformer pooling enables more flexible modeling of higher-order interactions among informative node representations.

Structural small-noise prior: A low-variance prior was introduced for the structural latent dimensions to constrain excessive freedom in these channels. This encourages the structural latent space to encode cleaner and more interpretable motif-related information, reducing the tendency to exploit unconstrained latent dimensions for shortcut solutions.

Single-block variational design: The architecture was simplified to a single variational latent block. Since the KL divergence term associated with the learnable GMM prior remains relatively small (due to beta typically ranging from 0.001 to 0.01), its regularization effect is limited but sufficient for this simplified setting, where posterior and prior can co-adapt more directly.

Channel separation for structural supervision: Dedicated latent channels were explicitly reserved for MMD-based structural supervision. This disentanglement prevents

Table 8: Per-seed classification accuracy (%) for Experiment 3 (Weighted Spurious Stress-Test) across random seeds and MMD weights γ. Mean $\pm$ sample standard deviation over the 8 seeds.

Seed	0.25	0.50	0.75	1.00	Baseline
2	100	68	83	97	88
12	71	99	95	98	93
22	90	97	92	97	98
32	97	87	97	97	99
42	87	96	85	98	70
52	99	97	98	96	96
62	62	97	97	99	85
72	99	97	99	98	99
Mean $\pm$ Std	88 ± 14	92 ± 10	93 ± 6	$\mathbf{98 \pm 1}$	91 ± 10

Table 9: Per-seed explanation ROC-AUC (%) for Experiment 3 (Weighted Spurious Stress-Test) across random seeds and MMD weights γ. Mean $\pm$ sample standard deviation over the 8 seeds.

Seed	0.25	0.50	0.75	1.00	Baseline
2	94	84	76	85	40
12	71	78	91	84	53
22	79	93	37	86	67
32	55	64	97	95	22
42	76	96	70	84	82
52	92	95	92	89	69
62	21	55	93	91	58
72	87	92	95	72	49
Mean $\pm$ Std	72 ± 24	82 ± 15	81 ± 20	$\mathbf{86 \pm 7}$	55 ± 19

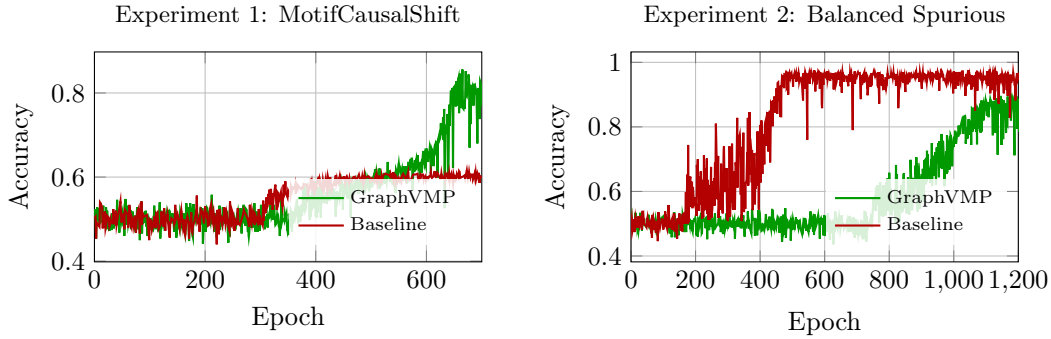

Figure 3: Accuracy curves for GraphVMP and the baseline on MotifCausalShift (Experiment 1, left) and Balanced Spurious (Experiment 2, right).

direct interference between structural alignment objectives and unconstrained latent encoding, allowing motif-specific information to be more effectively grounded.

C MotifCausalShift Dataset with Different Seeds

The results of graph classification accuracy and ROC-AUC for MotifCausalShift dataset with different seeds are shown in the tables below.

D Accuracy curves