INTERVENTION-VERIFIABLE HIERARCHICAL SELF-EXPLAINING HYPERGRAPH NEURAL NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Hypergraph neural networks effectively capture higher-order interactions, but identifying the structural basis of their predictions remains challenging. Post-hoc explainers attribute predictions to explanation subhypergraphs or hyperedge subsets after training without establishing whether the identified structure drives the output. To this end, we introduce Intervention-Verifiable Hierarchical Self-Explaining Hypergraph Neural Network (IV-SENN), a framework that enables direct verification of prediction changes under structural interventions on hypergraphs. IV-SENN represents higher-order hypergraph structure through h-motifs and captures recurring feature patterns using learnable feature prototypes. The additive concept-interaction head decomposes each class logit into linear, compositional, and quadratic interaction terms over the full concept encoding. With the trained head held fixed, IV-SENN exactly decomposes intervention-induced logit changes into structural and feature-linked contributions for counterfactual verification. Global explanations identify class-critical h-motifs from the learned structural weights of the additive concept-interaction head. Local explanations report signed structural and feature concept contributions to individual node predictions and ground influential structural concepts in concrete h-motif instances. Across four real-world datasets, IV-SENN remains competitive in node classification while attaining the highest Fid_KL and lowest SHD. On N-SYNTH, its estimated intervention effects achieve a Pearson correlation of 0.95 with the ground-truth changes in the predefined score.

1 INTRODUCTION

Many real-world systems involve interactions among groups of entities, such as co-authorship teams, biochemical complexes, and group communication. These higher-order interactions (HOIs) capture dependencies that cannot be fully represented by pairwise links (Battiston et al., 2020). Hypergraphs represent such interactions directly by allowing each hyperedge to connect an arbitrary number of nodes (Zhou et al., 2006). Building on this representation, hypergraph neural networks (HGNNs) extend neural message passing to hypergraph-structured data (Feng et al., 2019; Yadati et al., 2019) and have achieved better performance in node classification, recommendation, and drug discovery (Kim et al., 2024). Nevertheless, HGNNs remain difficult to interpret because the combinatorial space of higher-order structures obscures how specific hyperedges and node features contribute to individual predictions. We therefore study self-explaining HGNNs for node classification that expose the structural basis of each prediction and support intervention verification through controlled structural interventions.

GNN interpretability methods are commonly divided into post-hoc and self-explaining approaches. The former attribute predictions to subgraphs or counterfactual edits (Ying et al., 2019; Luo et al., 2020; Bajaj et al., 2021; Lucic et al., 2022; He et al., 2025), whereas the latter build explanations into the architecture (Alvarez-Melis & Jaakkola, 2018; Koh et al., 2020). Most graph-specific methods in both families operate on pairwise graph structures and use nodes, edges, or subgraphs as explanation units. Under clique expansion (Su et al., 2024), distinct hyperedges may induce the same graph edge, so an edit in the expanded graph need not map uniquely to the original hypergraph. This motivates explanation primitives defined directly on the hypergraph.

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.

Existing HGNN explainers remain predominantly post-hoc. HyperEX (Maleki et al., 2023) attributes predictions to node–hyperedge incidences, whereas CF-HyperGNNExplainer (Veglianti et al., 2026) identifies structural edits that change a prediction. Because explanation and prediction remain separate, the effect of an edit cannot be verified from the predictive computation itself. We call this requirement intervention consistency. It asks whether the output change produced by a structural edit can be recomputed from the model's predictive pathway.

Hypergraph motifs are recurrent higher-order patterns and established building blocks of local structure (Lee et al., 2020; Lotito et al., 2022; Lee et al., 2024; Benson et al., 2016). Their use in self-explaining prediction remains underexplored because hyperedges have indivisible semantics, h-motif associations grow combinatorially, and nodes can assume multiple roles. H-motifs capture higher-order structure in a hypergraph, but do not encode node features or distinguish the roles a node plays across overlapping motifs.

To address these challenges, we propose **IV-SENN**, the first self-explaining HGNN to incorporate h-motifs as structural concepts in its prediction path. IV-SENN represents higher-order structural patterns through h-motifs and captures intra-hyperedge feature patterns via matching scores from learnable feature prototypes, forming a node-level concept representation. We construct an additive concept-interaction head that maps this representation to class logits through linear, compositional, and interaction terms. Meanwhile, a separate h-motif-aware convolution supplies local h-motif role evidence without entering the prediction path. IV-SENN thereby supports global and local explanations, as well as counterfactual verification under structural edits. Our contributions are summarized as follows:

- We propose IV-SENN, a framework for intervention-verifiable self-explaining hypergraph neural networks. For the first time, IV-SENN incorporates h-motif statistics directly into the predictive pathway. We theoretically establish that class-logit changes induced by structural interventions can be exactly decomposed and verified at the prediction-head level.
- We construct a node-level concept representation that integrates higher-order structural patterns with intra-hyperedge feature patterns and leverage an additive concept-interaction head to capture synergistic effects among concepts. We further design an h-motif-aware convolution to localize the local structures associated with concept contributions.
- IV-SENN provides three complementary forms of explanation. Global explanations identify class-critical h-motifs, local explanations provide concept-level attributions and structural role localization, and counterfactual verification assesses whether the decomposed effects agree with the actual prediction changes induced by structural edits.
- We systematically evaluate IV-SENN on four real-world hypergraph datasets and construct N-SYNTH, a synthetic benchmark with observable ground truth for structural concepts and intervention effects. IV-SENN achieves the highest $\mathrm{Fid}_{\mathrm{KL}}$ and the lowest SHD on all four real-world datasets. On N-SYNTH, it achieves an h-motif recovery precision of 0.967 and a Pearson correlation of 0.948 between the estimated and ground-truth intervention effects.

2 Related Work

Self-Explaining Graph Neural Networks Self-explaining GNNs (SE-GNNs) couple prediction with explanation generation by identifying critical subgraphs through attention or masking mechanisms (Velickovic et al., 2018; Knyazev et al., 2019; Miao et al., 2022). Recent studies show that SE-GNN explanations can be redundant (Tai et al., 2025), misaligned with established notions of faithfulness (Azzolin et al., 2025a;b), or unrelated to the model's inference process (Azzolin et al., 2026), even when predictive performance remains strong. Other graph methods interpret predictions or representations through learned prototypes, graph-pattern analysis, grounded logical predicates, or reusable structural programs (Zhang et al., 2022; Shin et al., 2024; Wang et al., 2025; Geng et al., 2026; Jeon et al., 2026). These approaches remain designed for conventional pairwise graphs and do not extend directly to hyperedges that connect an arbitrary number of nodes.

Explainability for Hypergraphs Hypergraph explainability remains an emerging area, and most existing methods are post hoc. HyperEX (Maleki et al., 2023) assigns importance scores to node–hyperedge incidences and extracts explanation subhypergraphs using an attention-based framework.

SHypX (Su et al., 2024) uses discrete Gumbel–Softmax sampling for local explanation subhypergraphs and unsupervised concept extraction for global analysis. CF-HyperGNNExplainer (Veglianti et al., 2026) provides counterfactual explanations by identifying minimal structural edits that change the prediction. Self-explaining alternatives are less common. SHy (Yu et al., 2025) learns temporal phenotypes as personalized explanations. Graph-only methods such as SUNNY-GNN and G-NAMRFF require an intermediate graph representation and therefore cannot be applied directly to the original hypergraph. Most existing approaches explain predictions using node–hyperedge incidences, individual hyperedges, or local neighborhoods. They can identify important structures or find edits that change a prediction, but the effect of a structural edit on the output cannot usually be verified through the model's own prediction path.

3 Notation and h-motif Definitions

Hypergraph. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a hypergraph with node set $\mathcal{V}$ ($|\mathcal{V}| = N$) and hyperedge set $\mathcal{E}$ ($|\mathcal{E}| = M$). Hyperedge $e \in \mathcal{E}$ is a non-empty subset of $\mathcal{V}$. Node features are $\mathbf{X} \in \mathbb{R}^{N \times d}$. We focus on the task of node classification on hypergraphs, where each node v carries a label $y_v \in \{1, \ldots, C\}$.

H-motifs. Following Lee et al. (2020), an h-motif is defined as the overlap pattern of three connected hyperedges. This pattern is encoded by a 7-bit binary vector indicating which of the seven Venn regions are non-empty. After removing duplicates and disconnected configurations, there remain exactly 26 valid h-motifs, including closed types (hm1–16, hm23–26, all pairwise intersecting) and open types (hm17–22, containing a disjoint pair). We adopt the MoCHy enumeration and refer readers to that work for the complete taxonomy.

4 Method

4.1 Overview

IV-SENN maps a hypergraph $\mathcal{G}$ to node-level predictions and explanations. It concatenates an h-motif-based structural encoding with a learnable feature-prototype encoding to form the node representation $\mathbf{m}_v$ (Eq. 3). Each coordinate represents either an h-motif type or a feature prototype, and the additive concept-interaction head maps this representation to class logits through linear, compositional, and interaction terms. With the trained head held fixed across an intervention, the class-logit change induced by a structural edit can be computed exactly from the pre- and post-intervention encodings. A separate h-motif-aware convolution produces the class-agnostic role-evidence vector $\boldsymbol{\beta}_v$ for local explanation without contributing to the class logits. Figure 1 summarizes the architecture.

4.2 Structural Encoding

IV-SENN constructs h-motif-based structural encodings at two granularities. The node-level encoding enters the prediction path, while the hyperedge-level encoding supports local h-motif role evidence. For node v, let $c_{v,k}$ count its h-motif-k instances and $T_v = \sum_{k=1}^{26} c_{v,k}$. The structural encoding $\mathbf{m}_v^{\text{struct}} \in \mathbb{R}^{26}$ has entries $m_{v,k}^{\text{struct}} = c_{v,k}/T_v$ for $T_v > 0$ and zero otherwise.

At the hyperedge level, let $n_{e,k}$ be the number of h-motif-k instances containing e and $Z_e = \sum_{k=1}^{26} n_{e,k}$. We normalize the h-motif counts returned by MoCHy as follows:

$$\mathbf{t}_e^{\text{struct}} \in \mathbb{R}^{26}, \qquad t_{e,k}^{\text{struct}} = \begin{cases} n_{e,k}/Z_e, & Z_e > 0, \\ 0, & Z_e = 0, \end{cases} \tag{1}$$

The h-motif-aware convolution uses the structural match $s_k^{\text{struct}}(e) = \sigma(t_{e,k}^{\text{struct}})$. The sigmoid places the structural and feature matching scores on a common bounded scale while preserving the ordering of the normalized h-motif frequencies. Both structural encodings are computed from h-motif counts returned by MoCHy and remain fixed during training, so coordinate k retains the same h-motif semantics across datasets (Proposition 2).

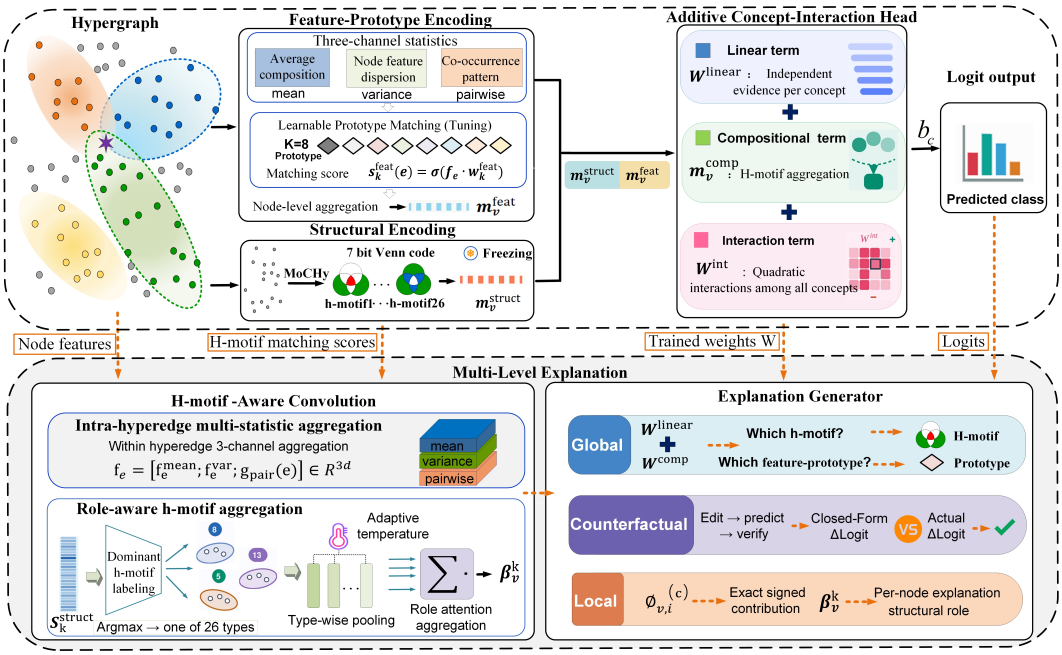


Figure 1: Overview of IV-SENN. H-motif structural concepts and learnable feature prototypes form $\mathbf{m}_v$ for prediction, while a separate h-motif-aware convolution supplies local h-motif role evidence.

4.3 Feature-Prototype Encoding

Structural concepts locate a hyperedge in the hypergraph, whereas feature prototypes characterize its composition. To retain information beyond the mean, we represent each hyperedge using three-channel statistics,

$$\mathbf{f}_e = [\mathbf{f}_e^{\text{mean}};\ \mathbf{f}_e^{\text{variance}};\ \mathbf{f}_e^{\text{pairwise}}] \in \mathbb{R}^{3d}, \tag{2}$$

where $\boldsymbol{\mu}_e = \mathbf{f}_e^{\text{mean}} = |e|^{-1}\sum_{v\in e}\mathbf{x}_v$, $\mathbf{f}_e^{\text{variance}} = |e|^{-1}\sum_{v\in e}(\mathbf{x}_v - \boldsymbol{\mu}_e)^{\odot 2}$, and, for $|e| \geq 2$, $\mathbf{f}_e^{\text{pairwise}} = \binom{|e|}{2}^{-1}\sum_{\{u,v\}\subset e}\mathbf{x}_u \odot \mathbf{x}_v$ (zero for $|e| = 1$). The pairwise channel averages the element-wise products of all node pairs in a hyperedge. It summarizes pairwise feature agreement but does not identify which nodes form distinct subgroups. K learnable prototypes $\mathbf{w}_k^{\text{feat}} \in \mathbb{R}^{3d}$ score each hyperedge by $s_k^{\text{feat}}(e) = \sigma\big(\mathbf{f}_e \cdot \mathbf{w}_k^{\text{feat}}\big)$. For node v, let $\mathcal{E}_v = \{e \in \mathcal{E} : v \in e\}$ denote its incident hyperedges. The node-level feature encoding $\mathbf{m}_v^{\text{feat}} \in \mathbb{R}^K$ averages the prototype matches, with $m_{v,k}^{\text{feat}} = |\mathcal{E}_v|^{-1}\sum_{e\in\mathcal{E}_v} s_k^{\text{feat}}(e)$ when $|\mathcal{E}_v| > 0$ and $m_{v,k}^{\text{feat}} = 0$ otherwise. Concatenating the structural and feature encodings gives the final node encoding

$$\mathbf{m}_v = [\mathbf{m}_v^{\text{struct}}; \mathbf{m}_v^{\text{feat}}] \in \mathbb{R}^{26+K}, \tag{3}$$

where K is the number of feature prototypes ($K = 8$ in our experiments). The highest-scoring hyperedges are retained as representative examples for interpreting each prototype; this labeling step does not affect prediction.

4.4 H-Motif-Aware Convolution

The h-motif-aware convolution localizes h-motif role evidence without contributing to the class logits. Within each hyperedge, multi-statistic aggregation summarizes feature variation beyond the mean. Across a node's incident hyperedges, role-aware h-motif aggregation distinguishes motif-dependent structural roles.

Intra-hyperedge multi-statistic aggregation. Standard message passing averages node features within a hyperedge, thereby obscuring differences among nodes. This intra-hyperedge multi-statistic aggregation mechanism reuses the three-channel statistics of the feature-prototype encoding (Eq. 2),

so the convolution could capture the mean, the variance, and the pairwise interactions, represented by element-wise products, of a hyperedge's node features rather than the mean alone.

Role-aware h-motif aggregation. A node may be incident to hyperedges associated with different h-motif types, placing it in distinct local structural contexts. Mean-based hypergraph convolution aggregates these hyperedges without preserving their motif-type distinctions. Within each incident hyperedge, a dual-match adaptive temperature $\tau(e;\tau_{\min})$ adjusts node attention according to $m(e) = \frac{1}{2}(\max_k s_k^{\text{struct}}(e) + \max_k s_k^{\text{feat}}(e))$. The lower bound $\tau_{\min}$ controls the strongest attainable sharpening and is set to 0.1 by default. Higher match strength lowers $\tau(e;\tau_{\min})$. For nonzero attention-logit gaps, this produces a sharper distribution, whereas larger temperatures avoid excessive concentration when the match is weak (Proposition 4). Each hyperedge is then assigned its dominant h-motif type, $\hat{k}_{\text{struct}}(e) = \arg\max_k s_k^{\text{struct}}(e)$. Hyperedges of the same type are pooled into $\mathbf{z}_v^k$, and the 26 type-level summaries are combined using learnable role attention:

$$\mathbf{h}_v^{(l+1)} = \sigma\Big(\mathbf{W}^{(l)} \sum_{k=1}^{26} \beta_v^k \, \mathbf{z}_v^k\Big). \tag{4}$$

The vector $\boldsymbol{\beta}_v \in \mathbb{R}^{26}$ satisfies $\beta_v^k \geq 0$ and $\sum_k \beta_v^k = 1$ and provides a nonnegative, class-agnostic distribution of structural-role evidence for the local explanation.

4.5 Additive Concept-Interaction Head

The Additive Concept-Interaction Head decomposes each logit into linear, compositional, and interaction terms.

Linear term. The linear term maps the full node encoding via $\mathbf{m}_v \cdot W_{:,c}^{\text{linear}}$, where $W^{\text{linear}} \in \mathbb{R}^{(26+K)\times C}$.

Compositional term. The compositional term combines the 26 h-motifs into functional units, $\mathbf{m}_v^{\text{comp}} = \mathbf{A}\mathbf{m}_v^{\text{struct}}$, where the attention matrix $\mathbf{A} \in \mathbb{R}^{K_{\text{comp}}\times 26}$ has entries

$$\alpha_{jk} = \text{softmax}_k\Big(\frac{\mathbf{q}_j^\top \mathbf{w}_k^{\text{struct}}}{\sqrt{d_{\text{comp}}}}\Big), \tag{5}$$

with learnable queries $\mathbf{q}_j$ and structural keys $\mathbf{w}_k^{\text{struct}}$, and $W^{\text{comp}} \in \mathbb{R}^{K_{\text{comp}}\times C}$ is the corresponding weight matrix. Proposition 5(a) shows that the input-independent $\mathbf{A}$ reparameterizes the linear term rather than enlarging the function class, while making h-motif compositions inspectable.

Interaction term. The interaction term captures reinforcement or suppression between concepts through $\sum_{i,j} m_{v,i} m_{v,j} W_{i,j,c}^{\text{int}}$. The resulting contribution is quadratic in the node encoding, where $W^{\text{int}} \in \mathbb{R}^{(26+K)\times(26+K)\times C}$ denotes the class-specific interaction tensor. For each class c, we enforce $W_{:,:,c}^{\text{int}} = (W_{:,:,c}^{\text{int}})^\top$, for example by parameterizing it as $(\Theta_c + \Theta_c^\top)/2$ (Proposition 6). Proposition 5(b) gives two classes that no linear classifier can separate but the interaction term can distinguish. The final logit is

$$\text{logit}_v^{(c)} = \underbrace{\mathbf{m}_v \cdot W_{:,c}^{\text{linear}}}_{\text{Linear term}} + \underbrace{\mathbf{m}_v^{\text{comp}} \cdot W_{:,c}^{\text{comp}}}_{\text{Compositional term}} + \underbrace{\sum_{i,j} m_{v,i} m_{v,j} W_{i,j,c}^{\text{int}}}_{\text{Interaction term}} + b_c. \tag{6}$$

Closed-form recomputation. Because the prediction is explicit in $\mathbf{m}_v$, an edit is evaluated by applying the unchanged head to the post-edit encoding. This exact head-level recomputation requires no parameter update or message passing, but excludes MoCHy recounting and feature re-encoding.

Training loss. The predictive pathway and feature prototypes are trained with

$$\mathcal{L}_{\text{pred}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{motif}}\mathcal{L}_{\text{motif}} + \lambda\big(\|W^{\text{linear}}\|_1 + \|W^{\text{comp}}\|_1 + \|W^{\text{int}}\|_1\big), \tag{7}$$

where $\mathcal{L}_{\text{CE}}$ is cross-entropy and $\mathcal{L}_{\text{motif}}$ combines prototype clustering and separation (Appendix C.3). The ℓ_1 terms encourage sparse head weights. The h-motif-aware convolution is fitted by the label-free objective in Appendix A.1, with stop-gradient preventing changes to the predictor or its guarantees.

4.6 MULTI-LEVEL EXPLANATION

IV-SENN provides global and local explanations and supports counterfactual verification under structural edits.

Global explanation. For class c, we combine the structural weights of the linear and compositional terms as $\widetilde{\mathbf{w}}_c^{\text{struct}} = W_{1:26,c}^{\text{linear}} + \mathbf{A}^\top W_{:,c}^{\text{comp}}$ and identify class-critical h-motifs using the selection rule in Appendix D. Structural concepts follow the predefined h-motif semantics, and feature prototypes are interpreted through representative hyperedges. Because quadratic interactions cannot be folded into the effective linear weights, their class-specific pairwise coefficients remain available for separate inspection and are excluded from the linear weight ranking. We evaluate the resulting global concepts through fidelity, sparsity, seed stability, feature–structure alignment, and planted-signal recovery in Sections 5.2–5.4.

Counterfactual verification. A global explanation readout summarizes the information encoded by the learned head weights. We intervene on the hypergraph and compare the predicted and observed changes. For a target node v and target h-motif-i, we rank the incident hyperedges participating in hm-i instances by their pre-edit target-type structural match $s_i^{\text{struct}}(e)$ and delete the lowest-ranked fraction $r \in \{0.3, 0.5\}$ individually. Full ranking and recounting details are provided in Appendix D.4. We evaluate the structural effect $\Delta\text{logit}^{\text{struct}}$ in closed form. On N-SYNTH, we apply the same intervention protocol and compare the fixed-head estimate with the corresponding change in the predefined ground-truth score. We also apply a matched null intervention at the same deletion ratio to assess whether the observed prediction change is specific to the planted signal h-motif or instead reflects nonspecific effects of structural deletion and h-motif spillover.

Local explanation. For class c, define the full effective linear weight $\mathbf{w}_c^{\text{eff}} = W_{:,c}^{\text{linear}} + [\mathbf{A}^\top W_{:,c}^{\text{comp}}; \mathbf{0}]$ and the signed contribution $\phi_{v,i}^{(c)} = m_{v,i} w_{i,c}^{\text{eff}} + m_{v,i}(W_{:,:,c}^{\text{int}} \mathbf{m}_v)_i$. Symmetry of the interaction matrix ensures that $\sum_i \phi_{v,i}^{(c)} = \text{logit}_v^{(c)} - b_c$, establishing additive completeness with respect to the head logit. A local explanation combines the contribution vector with the class-agnostic role distribution $\boldsymbol{\beta}_v$, the dominant structural and feature matches of the incident hyperedges, and representative hyperedges for the selected feature prototypes. Thus, ϕ quantifies how each concept supports or suppresses class c, whereas the h-motif-aware convolution indicates where the corresponding role evidence occurs. The deterministic grounding from class-critical h-motif types to local h-motif instances and their explanation subhypergraphs is given in Appendix D.1.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate IV-SENN and the baselines on four real-world hypergraph datasets(Cora, DBLP, Contact-high-school, and Contact-primary-school). Meanwhile, we use our synthetic dataset N-SYNTH to evaluate the accuracy of intervention-effect estimates. Dataset statistics and splits, along with the N-SYNTH generation protocol, are provided in Appendix C.

Baselines. We compare IV-SENN with the post-hoc methods HyperEX (Maleki et al., 2023), SHypX (Su et al., 2024), CF-HyperGNNExplainer (Veglianti et al., 2026) (abbreviated as CF-Hyper in Table 1), and the self-explaining methods SUNNY-GNN (Deng & Shen, 2024) and G-NAMRFF (Reddy et al., 2025). Further implementation details are provided in Appendix C.

Metrics. We report ACC and F1 for full-hypergraph classification, GSACC for classification on the retained h-motif explanation subhypergraph, Fid$^+$, Fid$^-$, and Fid_KL for fidelity, and SHD for stability. Experiment-specific metrics are defined in the corresponding sections and Appendix D.2.

5.2 PREDICTION PERFORMANCE AND EXPLANATION QUALITY

Table 1 compares predictive performance and explanation quality on the four real-world datasets. On Cora and DBLP, IV-SENN trails the strongest classification baseline by 1.8 and 4.9 ACC percentage

Table 1: Baseline experimental metrics on four real-world hypergraph datasets. Bold denotes the best result for each dataset. "–" denotes a non-comparable result. (Appendix D.2).

Method	Dataset	ACC	F1	GSACC	Fid$^-$	Fid_KL	SHD
HyperEX	Cora	0.7812	0.7698	0.7215	0.0613	0.8932	0.0315
	DBLP	0.8846	0.8721	0.8034	0.0924	0.6475	0.0287
	Contact-high-school	0.6538	0.6342	0.5863	0.0948	0.9128	0.0481
	Contact-primary-school	0.6440	0.6247	0.5775	0.1074	0.8991	0.0474
SHypX	Cora	**0.8118**	**0.8033**	0.7200	0.0950	0.9336	0.0273
	DBLP	**0.9085**	**0.9026**	0.7800	0.1300	0.6283	0.0172
	Contact-high-school	0.7812	0.7689	0.7451	0.0800	0.8386	0.0150
	Contact-primary-school	0.7643	0.7512	0.7287	0.0788	0.8260	0.0148
CF-Hyper	Cora	0.7735	0.7612	0.7418	–	–	0.0523
	DBLP	0.8721	0.8594	0.7021	–	–	0.0471
	Contact-high-school	0.6014	0.5876	0.5721	–	–	0.0648
	Contact-primary-school	0.5832	0.5687	0.5512	–	–	0.0638
SUNNY-GNN	Cora	0.7049	0.6879	0.7055	0.0749	0.6140	0.0275
	DBLP	0.7810	0.7747	0.7670	0.0534	0.5920	0.0275
	Contact-high-school	0.3658	0.3512	0.3547	0.2030	0.0971	0.0498
	Contact-primary-school	0.3421	0.3289	0.3312	0.2000	0.0956	0.0490
G-NAMRFF	Cora	0.6158	0.6351	0.6279	**0.0395**	0.7077	0.0881
	DBLP	0.8245	0.8149	**0.8609**	0.0579	0.7384	0.0674
	Contact-high-school	0.3385	0.3241	0.3298	0.2758	0.1554	0.0550
	Contact-primary-school	0.3124	0.2987	0.3012	0.2717	0.1531	0.0520
IV-SENN	Cora	0.7943	0.7816	**0.7582**	0.0483	**1.1734**	**0.0214**
	DBLP	0.8593	0.8472	0.8531	**0.0093**	**1.0371**	**0.0110**
	Contact-high-school	**0.9327**	**0.9351**	**0.9156**	**0.0285**	**1.3014**	**0.0108**
	Contact-primary-school	**0.8814**	**0.8382**	**0.8429**	**0.0512**	**1.2683**	**0.0096**

points, with a similar pattern in F1. On the two contact datasets, it leads the strongest baseline by 15.2 and 11.7 ACC percentage points. These results show that IV-SENN remains competitive on the citation and co-authorship datasets and performs best on both contact datasets.

Across all four datasets, IV-SENN achieves the highest Fid_KL and the lowest SHD. It ranks first in GSACC on three datasets and second on DBLP, trailing G-NAMRFF by 0.0078. Its Fid$^-$ is lowest on DBLP and the two contact datasets and second-lowest on Cora. Taken together, the high Fid_KL, low Fid$^-$, and low SHD indicate that the selected explanation subhypergraphs contain decision-critical information, largely preserve the prediction when retained, and remain stable across runs. Dataset-specific baseline behavior on the dense contact hypergraphs is discussed in Appendix D.2.

5.3 Feature–Structure Alignment

We evaluate h-motif-specific prototype alignment using Alignment@k on closed-type hm8 and open-type hm17 hyperedges (Appendix D.3). On N-SYNTH, prototypes 1–3 are enriched for hm8 and prototypes 4–5 for hm17, with reciprocal suppression on the other h-motif. The aligned prototypes attain scores of 1.43–1.72 for their associated h-motif type and 0.38–0.55 for the opposing type. Thus, five of the eight prototypes exhibit disjoint, h-motif-specific alignment beyond the empirical enrichment and suppression thresholds. Across nonzero structure–feature covariance levels, IV-SENN outperforms the unsupervised feature-clustering baseline, with the largest gap at an intermediate injection strength. Both methods remain at the background level when no covariance is introduced (Table 5). This analysis complements the intervention-verification setting in Section 5.4, which evaluates the fixed structural encoding under label-independent features.

The real-data analysis addresses a separate question and evaluates the sparsity and seed stability of the selected global structural concepts. Figure 2 shows sparse global h-motif sets on Cora, DBLP, and Contact-high-school. Across five training runs, the selected sets attain mean pairwise Jaccard similarities of 0.90–0.97 across the four real datasets, with a minimum observed value of 0.78. The

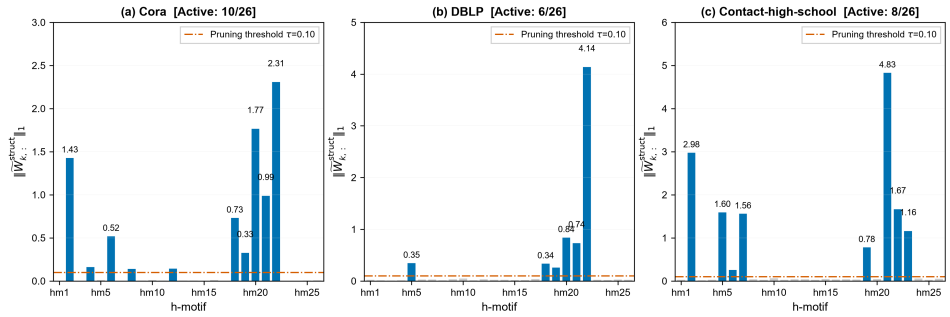


Figure 2: Effective h-motif subset on Cora, DBLP, and Contact-high-school: absolute effective structural linear weights per h-motif type.

Table 2: Planted h-motif recovery and head-level intervention verification on N-SYNTH (five-seed mean $\pm$ standard deviation; best in bold).

Condition	Precision ↑	CC_{dir} ↑	$CC_{Pearson}$ ↑	MAE ↓	rMAE ↓
IV-SENN (full)	**0.967±0.075**	**0.903±0.027**	**0.948±0.019**	**0.043±0.009**	**0.061±0.013**
w/o interaction term	0.933±0.091	0.861±0.043	0.902±0.034	0.067±0.014	0.095±0.020
w/o ℓ_1 sparsity	0.082±0.004	0.884±0.036	0.935±0.026	0.052±0.012	0.074±0.017
Feature confounding (30% PC)	0.900±0.091	0.842±0.049	0.887±0.039	0.071±0.018	0.101±0.026
Held-out nodes	0.883±0.108	0.826±0.057	0.874±0.046	0.078±0.021	0.112±0.031

corresponding result for Contact-primary-school and complete per-prototype and stability results are reported in Appendix D.3.

5.4 Intervention Verifiability

On N-SYNTH, signal h-motif precision evaluates whether the selected global structural concepts correspond to the planted hm8 and hm17 patterns. We also compare the structural logit changes computed with the trained head held fixed against the corresponding changes in the predefined ground-truth structural score (Table 2; metrics in Appendix D.2). The full model achieves a signal h-motif precision of 0.967 ± 0.075, $CC_{Pearson} = 0.948 \pm 0.019$, and $rMAE = 0.061 \pm 0.013$. These results show that the selected concepts closely match the planted structural signals and that the fixed-head estimates track their intervention effects. Removing the interaction term reduces $CC_{Pearson}$ to 0.902 ± 0.034 and increases rMAE to 0.095 ± 0.020. Removing ℓ_1 regularization produces a different failure mode. Signal h-motif precision falls from 0.967 to 0.082, while $CC_{Pearson}$ remains 0.935 and rMAE increases only to 0.074. This contrast indicates that sparsity primarily determines which structural concepts are selected, rather than the accuracy of the estimated change once an intervention is specified.

Under feature confounding and for held-out nodes, $CC_{Pearson}$ remains between 0.874 and 0.887, with rMAE between 0.101 and 0.112. The held-out results are weaker and more variable, providing evidence of node-level generalization within generated hypergraphs rather than cross-hypergraph generalization. Figure 3 shows a representative structural edit, while null-intervention controls and efficiency results are reported in Appendix D.4. Overall, the experiment provides controlled head-level verification rather than unrestricted causal identification.

5.5 Ablation Study

The ablation study evaluates each component of IV-SENN on DBLP (Table 3), complementing the controlled N-SYNTH intervention verification in Section 5.4. The largest losses arise from the higher-order interaction machinery. Replacing role-aware h-motif aggregation with mean pooling and removing the pairwise channel reduce GSACC by 17.71% and 12.45%, respectively. Both changes also increase Fid^{-} from 0.0093 to 0.1461 and 0.1585 and yield less concentrated explanations. Meanwhile, removing the compositional or interaction term and replacing the adaptive sched-

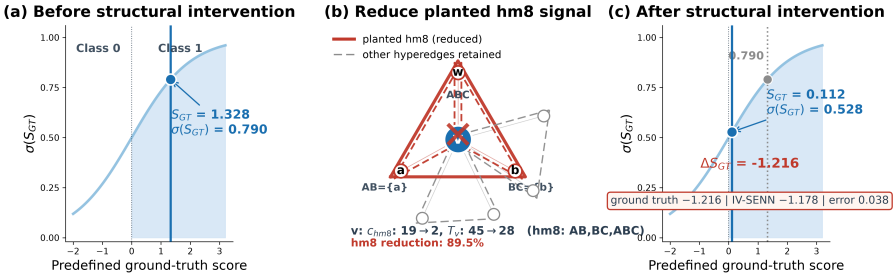


Figure 3: Sample-level structural intervention on N-SYNTH. An 89.5% reduction in the planted hm8 signal changes the predefined ground-truth score from 1.328 to 0.112, giving $\Delta S_{\mathrm{GT}} = -1.216$. IV-SENN estimates a logit change of -1.178, with an absolute error of 0.038.

Table 3: Ablation study on DBLP (five-seed means). Bold indicates the most favorable value in each metric column. Variant definitions are given in the table notes.

Variant	GSACC	Fid$^+$	Fid$^-$	Expl.Ent.	Eff.Mot.	WCS	Rel.Drop
Full	**0.8531**	**0.8286**	**0.0093**	**0.3273**	**6.5**	0.7398	—
w/o variance channel	0.8286	0.8041	0.0412	0.3441	7.2	**0.7512**	2.87%
w/o pairwise channel	0.7469	0.7592	0.1585	0.3888	8.5	0.6745	12.45%
mean pooling	0.7020	0.7020	0.1461	0.4257	11.0	—	17.71%
w/o compositional term	0.7696	0.7755	0.0898	0.3396	6.9	0.7156	9.79%
w/o interaction term	0.8039	0.7878	0.0416	0.3314	6.8	0.7435	5.77%
fixed $\tau(e) = 1$	0.8063	0.7878	0.0598	0.3604	8.2	0.6827	5.49%
w/o ℓ_1 regularization	0.8127	0.8249	0.0857	0.6833	25.0	0.6112	4.74%

The ablations remove the variance and pairwise channels, replace role-aware aggregation with mean pooling, remove the compositional term and interaction term, replace the adaptive schedule with the fixed temperature $\tau(e)$=1, and remove the ℓ_1 regularization, respectively.

ule with a fixed temperature produce smaller GSACC drops of 9.79%, 5.77%, and 5.49%. These results attribute the main explanation-quality gains to interaction modeling through role-sensitive structural aggregation and pairwise feature coupling rather than to a single linear component.

Removing ℓ_1 regularization reveals a complementary pattern. The GSACC drop is 4.74%, but the effective motif count increases from 6.5 to 25.0, explanation entropy more than doubles from 0.3273 to 0.6833, and WCS decreases from 0.7398 to 0.6112. This DBLP sparsity pattern is qualitatively consistent with the N-SYNTH intervention analysis, where signal h-motif precision falls from 0.967 to 0.082 without sparsity. Thus, sparsity primarily constrains which concepts carry the explanatory signal. The variance channel has the smallest GSACC effect (2.87%) and acts mainly as an explanation-fidelity refinement. Further sensitivity results appear in Appendix D.5.

6 Conclusion

We present IV-SENN, a self-explaining hypergraph neural network that makes prediction changes induced by structural edits verifiable within its own predictive computation. The model represents higher-order structure through h-motif-based structural encodings and recurring within-hyperedge feature patterns through learnable prototype matches. These representations connect class-level patterns with node-specific contributions and concrete local structures. The trained prediction head is held fixed across an intervention, and IV-SENN decomposes the resulting class-logit change under a structure-first attribution order. Experiments show that IV-SENN provides inspectable higher-order structural explanations and directly verifies the effects of structural edits on its predictions.

Limitations and future work. IV-SENN is currently limited to third-order h-motifs, and its feature prototypes are less informative when within-hyperedge features are weakly discriminative. Extending the structural vocabulary and reducing the cost of post-edit re-encoding are priorities for future work.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing, including improvements to grammar, clarity, and readability. They were not used to formulate the research questions, develop the method, write or execute code, conduct experiments, analyze results, or generate scientific claims. All AI-assisted edits were reviewed and revised by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs supporting the theoretical results are provided in Appendices A and B. Appendix C documents the N-SYNTH generation procedure, real-dataset preprocessing, implementation details, and hyperparameter settings. Appendix D defines the evaluation metrics and provides supplementary experimental protocols and analyses. The source code, data-generation scripts, and experimental configurations are available in an anonymous repository (link).

REFERENCES

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 7786–7795, 2018.

Steve Azzolin, Antonio Longa, Stefano Teso, and Andrea Passerini. Reconsidering faithfulness in regular, self-explainable, and domain invariant GNNs. In *International Conference on Learning Representations (ICLR)*, 2025a.

Steve Azzolin, Sagar Malhotra, Andrea Passerini, and Stefano Teso. Beyond topological self-explainable GNNs: A formal explainability perspective. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2045–2080, 2025b.

Steve Azzolin, Stefano Tesò, Bruno Lepri, Andrea Passerini, and Sagar Malhotra. GNN explanations that do not explain and how to find them. In *International Conference on Learning Representations (ICLR)*, 2026.

Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 5644–5655, 2021.

Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports*, 874:1–92, 2020. doi: 10.1016/j.physrep.2020.05.004.

Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016. doi: 10.1126/science.aad9029.

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are AllSet: A multiset function framework for hypergraph neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Jiale Deng and Yanyan Shen. Self-interpretable graph learning with sufficient and necessary explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):11749–11756, 2024. doi: 10.1609/aaai.v38i10.29059.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3558–3565, 2019. doi: 10.1609/aaai.v33i01.33013558.

Chuqin Geng, Ziyu Zhao, Zhaoyue Wang, Haolin Ye, Yuhe Jiang, and Xujie Si. LogicXGNN: Grounded logical rules for explaining graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2026.

Jesse He, Akbar Rafiey, Gal Mishne, and Yusu Wang. Explaining GNN explanations with edge gradients. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 884–895, 2025. doi: 10.1145/3711896.3736947.

Minseok Jeon, Seunghyun Park, and Jun-Gi Jang. ProgNet: Program-grounded evidence composition for interpretable graph classification. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 2002–2013, 2026. doi: 10.1145/3770855.3817844.

Sunwoo Kim, Soo Yong Lee, Yue Gao, Alessia Antelmi, Mirko Polato, and Kijung Shin. A survey on hypergraph neural networks: An in-depth and step-by-step guide. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 6534–6544, 2024. doi: 10.1145/3637528.3671457.

Boris Knyazev, Graham W Taylor, and Mohamed R Amer. Understanding attention and generalization in graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 4204–4214, 2019.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020.

Geon Lee, Jihoon Ko, and Kijung Shin. Hypergraph motifs: Concepts, algorithms, and discoveries. *Proceedings of the VLDB Endowment*, 13(11):2256–2269, 2020. doi: 10.14778/3407790.3407823.

Geon Lee, Seokbum Yoon, Jihoon Ko, Hyunju Kim, and Kijung Shin. Hypergraph motifs and their extensions beyond binary. *The VLDB Journal*, 33(3):625–665, 2024. doi: 10.1007/s00778-023-00827-8.

Quintino Francesco Lotito, Federico Musciotto, Alberto Montresor, and Federico Battiston. Higher-order motif analysis in hypergraphs. *Communications Physics*, 5(1):79, 2022. doi: 10.1038/s42005-022-00858-7.

Ana Lucic, Maartje A. ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. CF-GNNExplainer: Counterfactual explanations for graph neural networks. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4499–4511, 2022.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 19620–19631, 2020.

Sepideh Maleki, Ehsan Hajiramezanali, Gabriele Scalia, Tommaso Biancalani, and Kangway V. Chuang. Learning to explain hypergraph neural networks. In *ICML 2023 Workshop on TAG-ML*, 2023.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15524–15543, 2022.

Thummaluru Siddartha Reddy, Vempalli Naga Sai Saketh, and Mahesh Chandran. Interpretable and parameter efficient graph neural additive models with random fourier features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Sayan Saha and Sanghamitra Bandyopadhyay. Certified evaluation of model-level explanations for graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2026.

Yong-Min Shin, Sun-Woo Kim, and Won-Yong Shin. PAGE: Prototype-based model-level explanations for graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6559–6576, 2024. doi: 10.1109/TPAMI.2024.3379251.

Shiye Su, Iulia Duta, Lucie Charlotte Magister, and Pietro Liò. Explaining hypergraph neural networks: From local explanations to global concepts, 2024.

Wenxin Tai, Ting Zhong, Goce Trajcevski, and Fan Zhou. Redundancy undermines the trustworthiness of self-interpretable GNNs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58169–58188, 2025.

Fabiano Veglianti, Lorenzo Antonelli, and Gabriele Tolomei. Counterfactual explanations for hypergraph neural networks, 2026.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Xudong Wang, Ziheng Sun, Chris Ding, and Jicong Fan. Explainable graph representation learning via graph pattern analysis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3426–3434, 2025. doi: 10.24963/IJCAI.2025/381.

Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha P. Talukdar. HyperGCN: A new method of training graph convolutional networks on hypergraphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 1509–1520, 2019.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 9240–9251, 2019.

Leisheng Yu, Yanxiao Cai, Minxing Zhang, and Xia Hu. Self-explaining hypergraph neural networks for diagnosis prediction. In *Proceedings of the 6th Conference on Health, Inference, and Learning (CHIL)*, volume 287 of *Proceedings of Machine Learning Research*, pp. 908–924, 2025.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. ProtGNN: Towards self-explaining graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):9127–9135, 2022. doi: 10.1609/aaai.v36i8.20898.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 19, pp. 1601–1608, 2006.

A Theoretical Analysis

This appendix separates the theoretical guarantees from the empirical protocols used to evaluate them. We first formalize the distinction between head-derived decision contributions and convolutional role evidence, and then state the intervention, invariance, spillover, attention, expressiveness, and complexity results used by the method. Complete proofs are collected in Appendix B. Dataset construction and implementation details appear in Appendix C, followed by the evaluation definitions and additional experiments in Appendix D.

A.1 Decision Contributions and Local Role Evidence

Let $D = 26 + K$, $Q_c = W^{\text{int}}_{:,:,c} = Q_c^\top$, and

$$\mathbf{w}_c^{\text{eff}} = W^{\text{linear}}_{:,c} + [\mathbf{A}^\top W^{\text{comp}}_{:,c}; \mathbf{0}_K]. \tag{8}$$

For any node v, concept $i \in \{1, \dots, D\}$, and class c, define

$$\phi^{(c)}_{v,i} = m_{v,i} w^{\text{eff}}_{i,c} + m_{v,i}(Q_c \mathbf{m}_v)_i. \tag{9}$$

Summing Eq. 9 over i gives $\sum_i \phi^{(c)}_{v,i} = \text{logit}^{(c)}_v - b_c$. Symmetry assigns each off-diagonal interaction pair equally to its two endpoint concepts. This is a canonical accounting convention, rather than a unique causal allocation.

The convolutional role vector $\boldsymbol{\beta}_v$ has a different scope. It is nonnegative, class-agnostic evidence about the relative h-motif roles in the local neighborhood and is paired with $\boldsymbol{\phi}_v^{(c)}$ rather than substituted for it. The completeness identity applies only to the decision contributions. The role evidence is neither a unique hyperedge-level allocation nor a causal effect.

Label-free fitting of the h-motif-aware convolution branch. Let $\mathcal{V}_+ = \{v : \sum_{k=1}^{26} m_{v,k}^{\text{struct}} > 0\}$ and $\pi_{v,k} = m_{v,k}^{\text{struct}} / \sum_{j=1}^{26} m_{v,j}^{\text{struct}}$. The role parameters are fitted by

$$\mathcal{L}_{\text{role}} = \frac{1}{|\mathcal{V}_+|} \sum_{v \in \mathcal{V}_+} D_{\text{KL}}(\text{stopgrad}(\boldsymbol{\pi}_v) \| \boldsymbol{\beta}_v) \,. \tag{10}$$

Equation 10 updates only the h-motif-aware convolution parameters. All inputs shared with the concept encoder and predictor are detached, so this branch cannot alter any prediction parameter, concept value, prototype, composition matrix, interaction matrix, or intervention identity. Nodes outside $\mathcal{V}_+$ are assigned a uniform $\boldsymbol{\beta}_v$ only for reporting.

Theorem 1 (Exact Head-Level Intervention Decomposition)**.** *Fix a class c and trained parameters. Let $\mathbf{x}_0 = [\mathbf{s}; \mathbf{f}]$ be the pre-edit encoding, $\mathbf{x}_1 = [\mathbf{s} + \delta_s; \mathbf{f}]$, and $\mathbf{x}_2 = [\mathbf{s} + \delta_s; \mathbf{f} + \delta_f]$ the post-edit encoding, where $\mathbf{s} = \mathbf{m}_v^{\text{struct}}$ and $\mathbf{f} = \mathbf{m}_v^{\text{feat}}$. Define $F_c(\mathbf{x}) = \widetilde{\mathbf{w}}_c^\top \mathbf{x} + \mathbf{x}^\top Q_c \mathbf{x} + b_c$, with $\widetilde{\mathbf{w}}_c = W_{:,c}^{\text{linear}} + [\mathbf{A}^\top W_{:,c}^{\text{comp}}; \mathbf{0}]$ and $Q_c = W_{:,:,c}^{\text{int}} = Q_c^\top$. The logit difference therefore satisfies*

$$F_c(\mathbf{x}_2) - F_c(\mathbf{x}_0) = \underbrace{F_c(\mathbf{x}_1) - F_c(\mathbf{x}_0)}_{\Delta_c^{\text{struct}}} + \underbrace{F_c(\mathbf{x}_2) - F_c(\mathbf{x}_1)}_{\Delta_c^{\text{feat}}}, \tag{11}$$

and both terms are exactly computable from the endpoint encodings and the trained head. This ordered decomposition attributes structure–feature cross terms according to the stated structure-first path; reversing the path gives a different allocation but the same total change. Computing δ_s and δ_f after a hypergraph edit is outside this algebraic head-level guarantee.

Proposition 2 (Layer invariance of structural match scores)**.** *The structural match score $s_k^{\text{struct}}(e)$ is constant across training runs and parameter updates: since $\mathbf{t}_e^{\text{struct}}$ is precomputed and fixed by MoCHy, $s_k^{\text{struct}}(e)$ depends only on the hypergraph, not on learned weights.*

Proposition 3 (Renormalized node-level spillover bound)**.** *We consider a deletion-only intervention. For a given node u, let d_u denote the total number of destroyed h-motif instances across all types, and let $d_{u,j}$ denote the number of destroyed instances of type j. If $T_u > d_u$, then the renormalized change satisfies*

$$\Delta m_{u,j}^{\text{struct}} = \frac{m_{u,j}^{\text{struct}} d_u - d_{u,j}}{T_u - d_u}, \qquad \left|\Delta m_{u,j}^{\text{struct}}\right| \le \frac{m_{u,j}^{\text{struct}} d_u + d_{u,j}}{T_u - d_u}. \tag{12}$$

Consequently, for a target change $\eta = |\Delta m_{v,i}^{\text{struct}}| > 0$, the relative non-target spillover is at most $(m_{u,j}^{\text{struct}} d_u + d_{u,j})/((T_u - d_u)\eta)$. For deletion of one hyperedge e^*, $d_{u,j}$ is bounded by the number of pre-edit type-j h-motif instances that contain both u and e^*.

Adaptive temperature and attention. Attention in a hyperedge should concentrate on the nodes carrying the matched concept. With match strength $m(e) = \frac{1}{2}(\max_k s_k^{\text{struct}}(e) + \max_k s_k^{\text{feat}}(e))$, let $\tau_{\min} \in (0, 1]$ denote the minimum attention temperature, with the upper temperature fixed at one. The adaptive temperature is

$$\tau(e; \tau_{\min}) = \tau_{\min} + (1 - \tau_{\min})\big(1 - m(e)\big) \in (\tau_{\min}, 1], \tag{13}$$

where $\tau_{\min} = 0.1$ by default, recovering $\tau(e) = 0.1 + 0.9(1 - m(e))$. The lower endpoint is open because sigmoid match scores are strictly smaller than 1. For brevity, write $\tau_e = \tau(e; \tau_{\min})$. The raw attention logit $z_{vu} = \text{LeakyReLU}\big(\mathbf{a}^\top [\mathbf{h}_v; \mathbf{h}_u]\big)$ is normalized as

$$a_{vu} = \frac{\exp(z_{vu}/\tau_e)}{\sum_{w \in e} \exp(z_{vw}/\tau_e)}.$$

Let $\mathbf{a}_{v,e} = (a_{vu})_{u \in e}$ denote the resulting attention vector. A smaller temperature sharpens attention only when the raw logits have a nonzero gap.

Proposition 4 (Gap-dependent attention entropy bound)**.** *Let $n = |e| \geq 2$ and let $\delta = z_{(1)} - z_{(2)} \geq 0$ be the gap between the largest and second-largest raw attention logits. For Softmax temperature $\tau > 0$, define $q = [1 + (n-1)\exp(-\delta/\tau)]^{-1}$. The entropy then satisfies*

$$H(\mathbf{a}_{v,e} \mid \tau) \leq h_2(q) + (1-q)\log(n-1), \tag{14}$$

where $h_2(q) = -q\log q - (1-q)\log(1-q)$. For fixed nonzero gap the bound decreases as τ decreases and tends to zero only when $\delta/\tau \to \infty$; no deterministic-focus claim follows from match strength alone.

Proposition 5 (Compositional reparameterization and interaction necessity)**.** *Let $\mathbf{A}$ be the attention matrix defining the compositional term, and let W^{int} be the interaction tensor.*

(a) $\mathbf{A}$ is input-independent, so the compositional term $\mathbf{m}_v^{\text{comp}} = \mathbf{A}\,\mathbf{m}_v^{\text{struct}}$ is a linear reparameterization of $\mathbf{m}_v^{\text{struct}}$ that does not enlarge the function class of the linear term.

(b) There exist binary-label encodings that are not separable by any affine classifier but are separable by the interaction term. Hence, the interaction term strictly enlarges the linear function class.

Proposition 6 (Canonical symmetric interaction parameterization)**.** *For any square interaction matrix B_c and any encoding $\mathbf{m}$, $\mathbf{m}^\top B_c \mathbf{m} = \mathbf{m}^\top (B_c + B_c^\top)\mathbf{m}/2$. Hence the antisymmetric part is unobservable. Restricting $W^{\text{int}}_{:,:,c}$ to be symmetric leaves every prediction unchanged and removes this antisymmetric non-identifiability. This canonicalization does not imply full statistical identifiability when other design columns are collinear.*

Proposition 7 (Rademacher complexity)**.** *Conditional on a fixed sample of N node encodings, assume the following norm constraints: $\|W^{\text{linear}}\|_F \leq B_{\text{linear}}$, $\|W^{\text{comp}}\|_F \leq B_{\text{comp}}$, $\|W^{\text{int}}\|_F \leq B_{\text{int}}$, $\|\mathbf{m}_v\|_2 \leq R$, $\|\mathbf{m}_v^{\text{comp}}\|_2 \leq R_{\text{comp}}$, and $\|\mathbf{b}\|_2 \leq B_b$,*

$$\mathfrak{R}_N(\mathcal{H}) \leq \frac{\sqrt{C}}{\sqrt{N}}\big(RB_{\text{linear}} + R_{\text{comp}}B_{\text{comp}} + R^2 B_{\text{int}} + B_b\big). \tag{15}$$

B Proofs of Theorems and Propositions

B.1 Proof of Theorem 1

Write $D = 26 + K$, $\bar{\boldsymbol{\delta}}_s = [\delta_s; \mathbf{0}] \in \mathbb{R}^D$, and $\bar{\boldsymbol{\delta}}_f = [\mathbf{0}; \delta_f] \in \mathbb{R}^D$. Because the compositional map is input-independent, its linear contribution can be folded into $\widetilde{\mathbf{w}}_c$. Symmetry of Q_c gives, for any increment $\boldsymbol{\delta}$,

$$F_c(\mathbf{x} + \boldsymbol{\delta}) - F_c(\mathbf{x}) = \widetilde{\mathbf{w}}_c^\top \boldsymbol{\delta} + 2\mathbf{x}^\top Q_c \boldsymbol{\delta} + \boldsymbol{\delta}^\top Q_c \boldsymbol{\delta}. \tag{16}$$

Applying Eq. 16 first at $\mathbf{x}_0$ with $\bar{\boldsymbol{\delta}}_s$ and then at $\mathbf{x}_1$ with $\bar{\boldsymbol{\delta}}_f$ yields

$$\Delta_c^{\text{struct}} = \widetilde{\mathbf{w}}_c^\top \bar{\boldsymbol{\delta}}_s + 2\mathbf{x}_0^\top Q_c \bar{\boldsymbol{\delta}}_s + \bar{\boldsymbol{\delta}}_s^\top Q_c \bar{\boldsymbol{\delta}}_s, \tag{17}$$

$$\Delta_c^{\text{feat}} = \widetilde{\mathbf{w}}_c^\top \bar{\boldsymbol{\delta}}_f + 2\mathbf{x}_1^\top Q_c \bar{\boldsymbol{\delta}}_f + \bar{\boldsymbol{\delta}}_f^\top Q_c \bar{\boldsymbol{\delta}}_f. \tag{18}$$

The two differences telescope exactly to $F_c(\mathbf{x}_2) - F_c(\mathbf{x}_0)$. The cross term $2\bar{\boldsymbol{\delta}}_s^\top Q_c \bar{\boldsymbol{\delta}}_f$ enters the second step under the declared structure-first path. Reversing the order assigns it to the structural step instead, showing why the allocation is path-dependent although the total is invariant. Evaluation costs $O(D^2)$ after both endpoint encodings are available.

B.2 Proof of Proposition 2

$\mathbf{t}_e^{\text{struct}}$ is precomputed by MoCHy in the h-motif-based structural encoding and does not depend on the layer index l. Since $s_k^{\text{struct}}(e)$ applies the element-wise sigmoid to this fixed constant, it is layer-invariant. This makes $\hat{k}_{\text{struct}}(e) = \arg\max_k s_k^{\text{struct}}(e)$ consistent across layers, and the role-aware grouping $\mathcal{E}_v^k$ is fixed across layers.

B.3 Proof of Proposition 3

Before deletion, $c_{u,j} = m^{\text{struct}}_{u,j} T_u$. Afterwards the type-$j$ count and total count are $c'_{u,j} = c_{u,j} - d_{u,j}$ and $T'_u = T_u - d_u$. Therefore

$$m^{\text{struct}\prime}_{u,j} - m^{\text{struct}}_{u,j} = \frac{c_{u,j} - d_{u,j}}{T_u - d_u} - \frac{c_{u,j}}{T_u} = \frac{m^{\text{struct}}_{u,j} d_u - d_{u,j}}{T_u - d_u}.$$

The absolute bound follows from the triangle inequality, and division by $\eta > 0$ gives the relative bound. For a single deleted hyperedge, every destroyed type-j instance counted by $d_{u,j}$ must contain both this hyperedge and node u, which establishes the final statement.

B.4 Proof of Proposition 4

Let $p_{(1)}$ be the Softmax mass on the largest logit. Because every other logit is at most $z_{(1)} - \delta$,

$$p_{(1)} = \frac{1}{1 + \sum_{r=2}^{n} e^{(z_{(r)} - z_{(1)})/\tau}} \geq \frac{1}{1 + (n-1)e^{-\delta/\tau}} = q.$$

For fixed top mass p, entropy is maximized when the remaining mass is uniform, giving $h_2(p) + (1-p)\log(n-1)$. On $p \in [1/n, 1]$ this expression is non-increasing, and $p_{(1)} \geq q \geq 1/n$ yields the stated bound. The limit follows directly as $\delta/\tau \to \infty$.

B.5 Proof of Proposition 5

(a) By definition, $m^{\text{comp}}_{v,j} = \sum_{k=1}^{26} \alpha_{jk} m^{\text{struct}}_{v,k}$ with $\alpha_{jk} = \text{softmax}_k(\mathbf{q}_j^\top \mathbf{w}^{\text{struct}}_k / \sqrt{d_{\text{comp}}})$ depending only on trained parameters, not on input $\mathbf{m}_v$; hence $\mathbf{m}^{\text{comp}}_v = \mathbf{A}\,\mathbf{m}^{\text{struct}}_v$ for the constant matrix $\mathbf{A} = (\alpha_{jk})$, so the linear and compositional terms are linear in $\mathbf{m}^{\text{struct}}_v$ and do not enlarge the function class.

(b) Constructive proof. Consider two feature-encoding coordinates $m_1, m_2 \in \{0, 1\}$. Assign $(0, 0)$ and $(1, 1)$ to class A, and $(0, 1)$ and $(1, 0)$ to class B. The convex hulls of the two classes intersect at $(1/2, 1/2)$, so no affine hyperplane separates them. In contrast, $g(m_1, m_2) = (2m_1 - 1)(2m_2 - 1)$ is $+1$ on class A and -1 on class B. Since $g = 4m_1 m_2 - 2m_1 - 2m_2 + 1$, it is representable by the linear, interaction, and bias terms. Hence the interaction term strictly adds expressiveness.

B.6 Proof of Proposition 6

Decompose $B_c = S_c + A_c$, where $S_c = (B_c + B_c^\top)/2$ is symmetric and $A_c = (B_c - B_c^\top)/2$ is antisymmetric. For every real vector $\mathbf{m}$, the scalar $\mathbf{m}^\top A_c \mathbf{m}$ equals its own transpose, and $\mathbf{m}^\top A_c^\top \mathbf{m} = -\mathbf{m}^\top A_c \mathbf{m}$. Therefore, $\mathbf{m}^\top A_c \mathbf{m} = 0$. Thus B_c and S_c define the same function, whereas the antisymmetric parameters can never be recovered from predictions. Enforcing symmetry selects the unique representative of this equivalence class. It does not remove possible collinearity among the linear, compositional, self-interaction, and cross-interaction design columns, so no stronger least-squares uniqueness claim is made.

B.7 Proof of Proposition 7

Write the logit as an inner product: for class c, the logit is the sum of four blocks,

$$\text{logit}_c(\mathbf{m}_v) = \underbrace{\mathbf{m}_v \cdot W^{\text{linear}}_{:,c}}_{\text{block 1}} + \underbrace{\mathbf{m}^{\text{comp}}_v \cdot W^{\text{comp}}_{:,c}}_{\text{block 2}} + \underbrace{\left\langle \text{vec}(W^{\text{int}}_{:,:,c}), \text{vec}(\mathbf{m}_v \mathbf{m}_v^\top) \right\rangle}_{\text{block 3}} + b_c, \tag{19}$$

where the interaction term is rewritten via the Frobenius inner product. Let $\mathcal{H}_{\text{linear}}$, $\mathcal{H}_{\text{comp}}$, $\mathcal{H}_{\text{int}}$, and $\mathcal{H}_b$ be the function classes of the four blocks, with $\mathcal{H} = \bigoplus_{r=1}^{4} \mathcal{H}_r$. The Rademacher complexity is subadditive over the sum of function classes, so

$$\mathfrak{R}_N(\mathcal{H}) \leq \mathfrak{R}_N(\mathcal{H}_{\text{linear}}) + \mathfrak{R}_N(\mathcal{H}_{\text{comp}}) + \mathfrak{R}_N(\mathcal{H}_{\text{int}}) + \mathfrak{R}_N(\mathcal{H}_b). \tag{20}$$

Each block is a class of linear functions $\{x \mapsto \langle \theta, x \rangle : \|\theta\|_2 \leq B\}$ acting on inputs with $\|x\|_2 \leq R$; by the vector-contraction lemma and Khintchine inequality, its Rademacher complexity is at most

$\frac{1}{N}\mathbb{E}[\|\sum_{n=1}^{N}\sigma_n\boldsymbol{\phi}_r(\mathbf{m}_{v_n})\|_2]\,B_r \le \frac{R_r B_r}{\sqrt{N}}$. Applying this per block with $(R_1,B_1)=(R,B_{\text{linear}})$, $(R_2,B_2)=(R_{\text{comp}},B_{\text{comp}})$, $(R_3,B_3)=(R^2,B_{\text{int}})$, and $(R_4,B_4)=(1,B_b)$. Applying the standard vector-valued contraction bound to the C logits contributes the factor $\sqrt{C}$ and gives

$$\mathfrak{R}_N(\mathcal{H}) \le \frac{\sqrt{C}}{\sqrt{N}}\big(RB_{\text{linear}} + R_{\text{comp}}B_{\text{comp}} + R^2B_{\text{int}} + B_b\big). \tag{21}$$

The bound is not tight under structural dependence (nodes/hyperedges are not i.i.d.); its value is to show that ℓ_1 sparsity, by compressing the effective Frobenius norms, directly reduces the complexity term.

C Datasets and Implementation Details

C.1 N-SYNTH synthetic generation protocol

N-SYNTH contains 500 labeled nodes with 16-dimensional input features. For each of the five runs, we generate a stratified random split with 60% of the nodes for training, 20% for validation, and 20% for testing. The same split is used by all methods within a run.

Labels first determine h-motif-count distributions (rather than being caused by the generated hypergraph): $y_v \sim \text{Bernoulli}(0.5)$; $c_{v,\text{hm8}} \sim \text{Pois}(8)$ if $y_v = 1$ (else $\text{Pois}(2)$) and $c_{v,\text{hm17}} \sim \text{Pois}(2)$ if $y_v = 1$ (else $\text{Pois}(8)$). hm8/hm17 instances are planted in Venn patterns with v in the core region; noise instances (1–2 per irrelevant h-motif, label-independent) and background hyperedges (size 2–6, average degree $d^* = 8$–12) are added, and we retain nodes with $T_v \ge 30$ total instances so that the normalized frequencies $m_{v,k}^{\text{struct}} = c_{v,k}/T_v$ are stable. Under the Poisson means above, the expected total instance count is roughly $8 + 2 + 24 \cdot 1.5 \approx 46$ for both labels, so the $T_v \ge 30$ filter discards only the lower tail of the count distribution. Hyperedge reuse probability $p_{\text{reuse}} \in \{0.3, 0.5\}$ makes spillover real. Features $x_v \sim \mathcal{N}(0, I_d)$ are label-independent. MoCHy-E recounting keeps the relative discrepancy between the requested and recounted hm8/hm17 instance counts below 10%.

For intervention evaluation, we define the class-1 ground-truth score on the recounted structural encoding as

$$S_{\text{GT}}(\mathbf{m}_v^{\text{struct}}) = 3m_{v,\text{hm8}}^{\text{struct}} - 2m_{v,\text{hm17}}^{\text{struct}} + 8m_{v,\text{hm8}}^{\text{struct}}m_{v,\text{hm17}}^{\text{struct}}. \tag{22}$$

After a structural edit, MoCHy-E recounts the h-motif instances and renormalizes the structural encoding to obtain $\widetilde{\mathbf{m}}_v^{\text{struct}}$. The ground-truth score change is

$$\Delta S_{\text{GT}}(v) = S_{\text{GT}}(\widetilde{\mathbf{m}}_v^{\text{struct}}) - S_{\text{GT}}(\mathbf{m}_v^{\text{struct}}). \tag{23}$$

The coefficients are fixed before training and are not estimated from IV-SENN. This score is used only as an independent reference for intervention evaluation. It neither generates node labels nor enters model training.

C.2 Real datasets

Cora contains 2,708 nodes, 1,072 hyperedges, 1,433 input features, and seven classes. DBLP contains 41,302 nodes, 22,363 hyperedges, 1,425 input features, and six classes. Contact-high-school contains 327 nodes, 172,035 hyperedges, 16 input features, and nine classes, while Contact-primary-school contains 242 nodes, 106,879 hyperedges, one input feature, and 11 classes. Cora and DBLP use their standard node-classification splits. Because the two contact files do not provide masks, we use random 80/10/10 train/validation/test splits. All experiments use $K = 8$ learnable feature prototypes.

Baseline backbones. All baselines share the same splits and hyperparameter budgets. Each post-hoc baseline is applied with the backbone model of its original paper: AllSet (Chien et al., 2022) for HyperEX, AllSetTransformer for SHypX, and a three-layer HGCN with the propagation operator of Feng et al. (Feng et al., 2019) for CF-HyperGNNExplainer; the ACC/F1 of these rows therefore reflect each backbone's classification accuracy. SUNNY-GNN and G-NAMRFF are evaluated in their native self-interpretable and self-explaining architectures, respectively. The resulting comparison is therefore descriptive rather than a shared-backbone controlled experiment.

C.3 HYPERPARAMETERS

Prototype regularization. The prototype regularizer in Eq. 7 is

$$\mathcal{L}_{\text{motif}} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left\| \mathbf{f}_e - \mathbf{w}^{\text{feat}}_{k^*(e)} \right\|_2^2 + \frac{\gamma}{K(K-1)} \sum_{k \neq k'} \left[\max\left(0, \mu - \left\| \mathbf{w}^{\text{feat}}_k - \mathbf{w}^{\text{feat}}_{k'} \right\|_2 \right) \right]^2, \quad (24)$$

where $k^*(e) = \arg\max_k s^{\text{feat}}_k(e)$, $\gamma > 0$, and $\mu > 0$ is the separation margin. The first term pulls each hyperedge toward its assigned prototype, while the bounded margin discourages prototype collapse without driving prototype norms to infinity. We set $\mu = 1.0$ and $\gamma = 1.0$.

The additive concept-interaction head uses $W^{\text{linear}} \in \mathbb{R}^{(26+8)\times C}$, $K_{\text{comp}} = 8$, $W^{\text{int}} \in \mathbb{R}^{34\times 34\times C}$. The h-motif-aware convolution has two layers with a hidden dimension of 64. Role-attention logits are obtained from a single linear map from 26 inputs to 26 outputs, followed by a softmax. IV-SENN is trained with Adam at a learning rate of 0.01 for at most 500 epochs. We apply early stopping with a patience of 50 based on the validation loss. Weight decay is selected per dataset. We set $\lambda_{\text{motif}} = 0.1$ and $\lambda = 10^{-3}$ for the linear term, compositional term, and interaction term weights (set to 0 in the ℓ_1-removal ablation). All aggregate experiments use five seeds. Results shown with $\pm$ entries report the cross-seed mean and standard deviation; aggregate tables and sensitivity curves containing single values report the five-seed mean unless stated otherwise. The local case study is a single-checkpoint illustration rather than a cross-seed aggregate. All experiments were conducted on a single NVIDIA RTX 3090 GPU with 24 GB of GPU memory.

D ADDITIONAL EXPERIMENTS

D.1 H-MOTIF-GROUNDED EXPLANATION SUBHYPERGRAPHS

For a target node v with predicted class $c = \widehat{y}_v$, let $\mathcal{M}_{v,k}$ denote the set of enumerated h-motif-k instances that contain v. Each instance $\mu = (e_1, e_2, e_3)$ consists of three hyperedges, and $k(\mu)$ denotes its h-motif type. We restrict the candidate instances to the class-critical types selected by Eq. 37:

$$\mathcal{M}^{(c)}_v = \bigcup_{k \in \mathcal{K}_c} \mathcal{M}_{v,k}. \quad (25)$$

To distinguish multiple local instances of the same type without introducing an additional trainable prediction component, we define the instance-level match

$$L_v(\mu) = \frac{1}{3} \sum_{e \in \mu} \frac{1}{2} \left(s^{\text{struct}}_{k(\mu)}(e) + \max_j s^{\text{feat}}_j(e) \right). \quad (26)$$

The grounding score combines the class-specific structural contribution, the class-agnostic role evidence, and the instance-level match:

$$J^{(c)}_{v,\mu} = \mathbf{1}[k(\mu) \in \mathcal{K}_c] \, [\phi^{(c)}_{v,k(\mu)}]_+ \, \beta_{v,k(\mu)} \, L_v(\mu), \qquad [x]_+ = \max(x, 0). \quad (27)$$

We select $\mu^*_v = \arg\max_{\mu \in \mathcal{M}^{(c)}_v} J^{(c)}_{v,\mu}$ and define its three constituent hyperedges as the explanation subhypergraph, $S_v = \{e : e \in \mu^*_v\}$. Thus, the unified per-node budget $K_{\text{exp}} = 3$ retains one complete h-motif instance rather than three independently ranked hyperedges. If all positive structural contributions in the candidate set are zero, $|\phi^{(c)}_{v,k(\mu)}|$ replaces $[\phi^{(c)}_{v,k(\mu)}]_+$ in Eq. 27. If no candidate instance exists, we replace $\mathcal{K}_c$ in Eqs. 25–27 by the singleton containing the highest-contribution structural type that has a local instance; nodes with no local h-motif instance receive the empty structural explanation. In the mean-pooling ablation, the unavailable learned role distribution is replaced by a uniform distribution over the h-motif types observed locally at v. The incidence structure, structural counts, and feature-prototype averages are recomputed on the retained or removed subhypergraph before evaluation, while all learned parameters remain fixed. This grounding procedure affects explanation selection only and does not enter the full-hypergraph prediction path.

D.2 METRICS

For prediction we report ACC and Macro-F1 (F1). For C classes, F1 is

$$\mathrm{F1} = \frac{1}{C}\sum_{c=1}^{C}\frac{2\,\mathrm{Prec}_c\,\mathrm{Rec}_c}{\mathrm{Prec}_c + \mathrm{Rec}_c}, \tag{28}$$

where the classwise term is set to zero when its denominator is zero. Recent work has shown that class score alone does not fully characterize the quality of model-level explanations (Saha & Bandyopadhyay, 2026). For explanations we report ground-truth subgraph accuracy (GSACC), Fidelity+ (Fid$^+$), Fidelity− (Fid$^-$), Fid_KL, and structural Hamming distance (SHD), computed under a unified per-node top-K_{exp} explanation budget ($K_{\mathrm{exp}} = 3$) across all methods. Because Table 1 uses each method's native backbone and explanation unit, the resulting cross-method metrics are descriptive rather than a shared-backbone controlled comparison. **Fid_KL.** This metric is the KL divergence between the full-hypergraph predictive distribution and that after removing the explanation subhypergraph. A larger value indicates that the removed structure carries more decision-critical information. Fidelity+ (Fid$^+$) is the fraction of samples whose prediction changes after the explanation subhypergraph is removed; higher is better. Fidelity− (Fid$^-$) is the fraction of samples whose prediction changes when only the explanation subhypergraph is retained; lower is better. Writing p_v for the full predictive distribution, p_v^{-S} for the distribution after removing explanation S_v, and p_v^{S} for the distribution retaining only S_v, these metrics are

$$\mathrm{Fid_KL} = N^{-1}\sum_{v} D_{\mathrm{KL}}(p_v \| p_v^{-S}), \tag{29}$$

$$\mathrm{Fid}^+ = N^{-1}\sum_{v}\mathbf{1}\left[\arg\max_c p_{v,c}^{-S} \neq \arg\max_c p_{v,c}\right], \tag{30}$$

$$\mathrm{Fid}^- = N^{-1}\sum_{v}\mathbf{1}\left[\arg\max_c p_{v,c}^{S} \neq \arg\max_c p_{v,c}\right]. \tag{31}$$

GSACC is the classification accuracy obtained when only the selected explanation subhypergraph is retained:

$$\mathrm{GSACC} = N^{-1}\sum_{v}\mathbf{1}\left[\arg\max_c p_{v,c}^{S} = y_v\right]. \tag{32}$$

Unlike Fid$^-$, which measures agreement with the full-hypergraph prediction, GSACC compares the retained-subhypergraph prediction with the ground-truth label. For the DBLP ablation in Table 3, relative GSACC drop is computed as $\mathrm{Rel.Drop} = (\mathrm{GSACC}_{\mathrm{full}} - \mathrm{GSACC}_{\mathrm{variant}})/\mathrm{GSACC}_{\mathrm{full}} \times 100\%$, with $\mathrm{GSACC}_{\mathrm{full}} = 0.8531$. SHD is the pairwise normalized Hamming distance between binary explanation masks across seeds. Let $q_c = |\widetilde{\mathbf{w}}_c^{\mathrm{struct}}|/\|\widetilde{\mathbf{w}}_c^{\mathrm{struct}}\|_1$; the Weighted Role-Consistency Score is $\mathrm{WCS}_v = 1 - \frac{1}{2}\|\boldsymbol{\beta}_v - q_{\widehat{y}_v}\|_1$. WCS measures role-evidence alignment with class-level h-motif semantics; it is not a fidelity or contribution-decomposition identity.

For the concentration diagnostics in Table 3 and Figure 6, we use a pre-selection grounded motif-importance vector rather than the prediction-head weights or the role distribution alone. For test node v, predicted class $c = \widehat{y}_v$, and structural type $k \in \{1, \ldots, 26\}$, let

$$\ell_{v,k} = \begin{cases} \max_{\mu\in\mathcal{M}_{v,k}} L_v(\mu), & \mathcal{M}_{v,k} \neq \varnothing, \\ 0, & \mathcal{M}_{v,k} = \varnothing, \end{cases} \qquad a_{v,k} = |\phi_{v,k}^{(c)}|\,\beta_{v,k}\,\ell_{v,k}. \tag{33}$$

Thus, $a_v \in \mathbb{R}^{26}_{\geq 0}$ is the unsigned counterpart of the instance-grounding score: it combines decision strength, local role evidence, and the best available instance match, while retaining both supporting and suppressing motifs. It is used only as a diagnostic and does not enter training, prediction, intervention decomposition, or final subhypergraph selection.

Let $\mathcal{V}_a = \{v \in \mathcal{V}_{\mathrm{test}} : \|a_v\|_1 > 0\}$ and $p_{v,k} = a_{v,k}/\|a_v\|_1$. Explanation entropy and effective motif count are

$$\mathrm{Expl.Ent.} = \frac{1}{|\mathcal{V}_a|}\sum_{v\in\mathcal{V}_a}\frac{-\sum_{k=1}^{26} p_{v,k}\log p_{v,k}}{\log 26}, \tag{34}$$

$$\mathrm{Eff.Mot.} = \frac{1}{|\mathcal{V}_a|}\sum_{v\in\mathcal{V}_a}\frac{\|a_v\|_1^2}{\|a_v\|_2^2}. \tag{35}$$

Both are computed per seed and then averaged over the five seeds. The associated explanation coverage is $|\mathcal{V}_a|/|\mathcal{V}_{\text{test}}|$. Zero-importance nodes are excluded from the concentration averages so that missing explanations are not mistaken for concentrated ones. Lower Expl.Ent. and Eff.Mot. indicate greater concentration conditional on nonzero coverage and adequate fidelity; neither is itself a fidelity or causal metric.

By contrast, the active motif counts displayed in Figures 2 and 4 are hard global counts at $\tau_w = 0.10$, based on the cross-class effective head-weight magnitude. They are distinct from the node-level soft effective motif count above. CF-HyperGNNExplainer's explanation is a minimal counterfactual edit set whose removal is intended to flip the prediction, so its fidelity metrics are marked "–" in Table 1 rather than treated as directly comparable with attribution-based explanations. For ground-truth score changes o_v and model-predicted changes d_v, $\text{CC}_{\text{dir}} = N^{-1}\sum_v \mathbf{1}[\text{sign}(d_v) = \text{sign}(o_v)]$, $\text{CC}_{\text{Pearson}}$ is their Pearson correlation, and $\text{MAE} = N^{-1}\sum_v |d_v - o_v|$. Signal h-motif precision compares the selected structural h-motif set with $\{\text{hm8}, \text{hm17}\}$. For $\sum_v |o_v| > 0$, relative MAE is defined as

$$\text{rMAE} = \frac{\sum_v |d_v - o_v|}{\sum_v |o_v|} = \frac{\text{MAE}}{N^{-1}\sum_v |o_v|}. \tag{36}$$

The denominator is the MAE of a predictor that always predicts zero change. Thus, $\text{rMAE} < 1$ indicates improvement over this baseline, and lower values indicate smaller errors relative to the intervention scale. We compute rMAE separately for each seed before reporting the cross-seed mean and standard deviation. The Feature-Linkage Ratio is $\text{FLR}_v = |\Delta_v^{\text{feat}}|/(|\Delta_v^{\text{struct}}| + |\Delta_v^{\text{feat}}| + \epsilon)$ and $\text{SR}_v = \max_{u,j\neq i} |\Delta m_{u,j}^{\text{struct}}|/(|\Delta m_{v,i}^{\text{struct}}| + \epsilon)$. Alignment@k is defined in Appendix D.3.

Global explanation selection and contribution. For class c, define $\widetilde{\mathbf{w}}_c^{\text{struct}} = W_{1:26,c}^{\text{linear}} + \mathbf{A}^\top W_{:,c}^{\text{comp}}$. The class-critical h-motifs are

$$\mathcal{K}_c = \{k : |\widetilde{w}_{k,c}^{\text{struct}}| > \tau_w\} \cup \{k : k \in \text{Top-}K_{\text{top}} \text{ of } |\widetilde{\mathbf{w}}_c^{\text{struct}}|\}, \tag{37}$$

and the three logit terms are

$$\text{Contrib}^{\text{linear}} = \sum_k m_{v,k} W_{k,c}^{\text{linear}}, \tag{38}$$

$$\text{Contrib}^{\text{comp}} = \sum_j m_{v,j}^{\text{comp}} W_{j,c}^{\text{comp}}, \tag{39}$$

$$\text{Contrib}^{\text{int}} = \sum_{i,j} m_{v,i} m_{v,j} W_{i,j,c}^{\text{int}}, \tag{40}$$

each reported as $|\text{Contrib}^{(r)}|/\sum_{r'\in\{\text{linear,comp,int}\}} |\text{Contrib}^{(r')}|$ (excluding the bias b_c). The largest symmetric interaction pairs can be inspected separately and are not folded into the effective linear-weight ranking.

Baseline artifacts on dense hypergraphs. Two baselines display characteristic artifacts. SUNNY-GNN's learned subgraphs are the least decision-critical (Fid_KL 0.10–0.61; ≈ 0.1 on the dense contact graphs, where removing them barely moves the prediction), and HyperEX's attention degenerates on high-degree contact nodes (Fid_KL ≈ 0.9 with only moderate GSACC ≈ 0.58); G-NAMRFF shows the same near-degenerate selection on contact data (GSACC $\approx$ ACC, low Fid_KL). These patterns vanish on the low-degree co-authorship graphs and are consistent with attention or feature degeneration under heavy-tailed hyperedge degree. The present comparison does not isolate the degree distribution as a causal explanation for this behavior.

D.3 Feature–Structure Alignment: Metric and Verification

Alignment@k (mean-score form). Let $s_k^{\text{feat}}(e) = \sigma(\mathbf{f}_e \cdot \mathbf{w}_k^{\text{feat}})$ be the matching score of hyperedge e to prototype k, where $\mathbf{f}_e$ is the three-channel representation in Eq. 2. The alignment of prototype k toward h-motif type $m \in \{\text{hm8}, \text{hm17}\}$ is the ratio of mean matching scores,

$$\text{Alignment@k}(m) = \frac{\mathbb{E}[\, s_k^{\text{feat}}(e) \mid e \text{ has dominant h-motif type } m\,]}{\mathbb{E}[\, s_k^{\text{feat}}(e) \mid e \in \mathcal{E}\,]}.$$

Table 4: Feature–structure alignment on N-SYNTH. Alignment@k > 1.4 or < 0.6 indicates significant enrichment or suppression of prototype k in the corresponding h-motif type (well beyond the dispersion of the background prototypes, $\approx 1.0 \pm 0.08$); values are 5-seed mean $\pm$ std.

Prototype k	Dominant type	Alignment@k(hm8)	Alignment@k(hm17)	Semantics
1	hm8	**1.72±0.12**	0.42±0.08	high closed density
2	hm8	**1.58±0.10**	0.51±0.07	closed + high feat.
3	hm8	**1.49±0.09**	0.55±0.06	closed (weak)
4	hm17	0.38±0.06	**1.66±0.11**	open structure
5	hm17	0.52±0.07	**1.43±0.09**	open + low feat.
6	mixed	1.02±0.08	0.98±0.07	background
7	mixed	0.95±0.07	1.05±0.08	background
8	mixed	1.08±0.09	0.92±0.07	background

The dominant type is $\widehat{k}(e) = \arg\max_k n_{e,k}$, equivalently $\arg\max_k t^{\text{struct}}_{e,k}$. Exact ties are resolved by a fixed hash of the hyperedge and motif identifiers that is independent of node features and labels. Hyperedges with $Z_e = 0$ are treated as background and excluded from the hm8/hm17 conditional alignment statistics. Alignment $= 1$ means the prototype matches type-m hyperedges exactly at the background rate; > 1 indicates enrichment. The threshold (1.4 / 0.6) is set from the dispersion of the background prototypes (1.0 ± 0.08), so that no background prototype crosses it.

Covariant N-SYNTH protocol for feature–structure alignment. Labels $y_v \sim \text{Bernoulli}(0.5)$; planted signal instances $c_{v,\text{hm8}} \sim \text{Pois}(8 \mid y{=}1)$, $\text{Pois}(2 \mid y{=}0)$ (and the reverse for hm17); hm8-type hyperedges draw members from a positive node pool and hm17-type hyperedges from a negative node pool, plus noise and background hyperedges. Before hyperedge construction, define the generation-stage affinities

$$m_{\text{pos},v} = \frac{c_{v,\text{hm8}}}{c_{v,\text{hm8}} + c_{v,\text{hm17}}}, \qquad m_{\text{neg},v} = \frac{c_{v,\text{hm17}}}{c_{v,\text{hm8}} + c_{v,\text{hm17}}}, \tag{41}$$

with both values set to $1/2$ when the denominator is zero. The pools are $\mathcal{P}_{\text{pos}} = \{v : m_{\text{pos},v} > m_{\text{neg},v}\}$ and $\mathcal{P}_{\text{neg}} = \{v : m_{\text{neg},v} > m_{\text{pos},v}\}$. Affinity ties are assigned by a fixed hash of the node identifier, independent of features and labels. Let

$$\Delta = \mathbb{E}[m_{\text{pos},v} \mid v \in \mathcal{P}_{\text{pos}}] - \mathbb{E}[m_{\text{pos},v} \mid v \in \mathcal{P}_{\text{neg}}]. \tag{42}$$

Node features are $x_v = \epsilon_v + \alpha m_{\text{pos},v} u$, where $\epsilon_v \sim \mathcal{N}(0, I_d)$, $\|u\|_2 = 1$, and u is sampled once and kept fixed and hidden from the model. Thus, hm17 receives no separate negative injection: its relative suppression along u follows from the lower m_{pos} values in $\mathcal{P}_{\text{neg}}$, with the between-pool projected-mean gap equal to $\alpha\Delta$. Unlike the intervention-verification setting in Section 5.4, which uses independent features, this protocol injects structure–feature covariance so that prototypes have structure to bind to; the two variants answer complementary questions (structure-only recovery vs. feature–structure binding).

Mechanistic interpretation of the alignment trends. Writing $M = m_{\text{pos},v}$ and $p = \Pr(v \in \mathcal{P}_{\text{pos}})$, independence of M and ϵ_v gives the following exact node-level identity under injection strength α:

$$\text{Cov}(x_v) - I_d = \alpha^2 \,\text{Var}(M) u u^\top, \tag{43}$$

where

$$\text{Var}(M) = p\,\text{Var}(M \mid \mathcal{P}_{\text{pos}}) + (1-p)\,\text{Var}(M \mid \mathcal{P}_{\text{neg}}) + p(1-p)\Delta^2. \tag{44}$$

Hence $\alpha^2 p(1-p)\Delta^2 u u^\top$ is the between-pool component, not the complete perturbation unless the affinity is constant within each pool. Under independent isotropic noise, the hyperedge-mean channel satisfies

$$\text{Cov}(\boldsymbol{\mu}_e) - \mathbb{E}[|e|^{-1}] I_d = \alpha^2 \,\text{Var}(\overline{M}_e) u u^\top, \qquad \overline{M}_e = |e|^{-1} \sum_{v \in e} M_v. \tag{45}$$

The coefficient therefore depends on hyperedge size, member composition, and reuse. Because the variance and pairwise channels in Eq. 2 are nonlinear, we use this calculation to characterize

Table 5: Average Alignment in the hm8 direction (max over the eight prototypes) under increasing structure–feature covariance α on N-SYNTH (Section 5.3). At $\alpha = 2$, the IV-SENN value matches the maximum prototype alignment reported in Table 4. The gap peaks at intermediate α and vanishes at $\alpha = 0$ (no covariance, no binding for any method).

α	0	0.3	1	2	4	6
k-means (unsupervised baseline)	1.00	1.01	1.09	1.31	1.82	2.19
IV-SENN (max prototype)	1.00	1.16	1.45	1.72	2.04	2.22
gap (IV-SENN − k-means)	0.00	0.15	0.36	**0.41**	0.22	0.03

the quadratic weak-signal scaling of the mean-channel covariance rather than assert the same exact coefficient for the complete three-channel representation. The k-means baseline uses only these hyperedge-feature statistics, whereas the supervised prototypes additionally receive the task loss. Consistent with this difference, the measured $\mathrm{gap}(\alpha) = \mathrm{Alignment}^{\mathrm{IVSENN}} - \mathrm{Alignment}^{\mathrm{kmeans}}$ is positive for nonzero α, largest at intermediate injection, and zero at $\alpha = 0$. A complementary inhibition pattern is also observed: the sets of hm8- and hm17-enriched prototypes are disjoint in the reported runs, and each enriched prototype has an alignment below one for the opposing type. This pattern is consistent with hm8 and hm17 being mutually exclusive dominant types. Without a structure-input ablation, this comparison supports predictive feature–structure association but does not identify a causal contribution of the h-motif-based structural encoding.

Uncertainty and dependence. Table 4 reports means and standard deviations over five seeds. The thresholds of 1.4 and 0.6 are empirical effect-size criteria based on the dispersion of the background prototypes; they are not formal significance thresholds. A permutation test could compare the target-type fraction among a prototype's top-matching hyperedges with its global frequency, but no such per-prototype p-values are reported here. Moreover, the top-matching hyperedge sets may overlap, so the eight prototypes are not independent replicates. The number of aligned prototypes should therefore be interpreted descriptively.

Relation to intervention verification. Section 5.4 shows that the h-motif-based structural encoding recovers the planted hm8/hm17 h-motifs under feature-independent data, with full-model precision 0.967 ± 0.075. The feature–structure alignment analysis in Section 5.3 shows that, under covariant data, the feature prototypes bind the same two types beyond the unsupervised feature-statistics baseline. The two experiments use different feature regimes by design and are complementary (structure-only recovery vs. feature–structure binding); they are not a single evidential chain, and the explanation-quality metrics in Section 5.2 are not used as evidence for prototype–structure alignment.

At the displayed threshold of $\tau_w = 0.10$, Contact-primary-school retains 9 of the 26 structural h-motif dimensions. The largest retained weights occur for hm21 (4.26), hm2 (3.46), and hm5 (1.91), while the remaining active types have substantially smaller magnitudes. This concentration supports a sparse linear structural read-out, but it does not summarize the interaction term.

The selected h-motif sets are stable across the five runs. Mean pairwise Jaccard similarity is 0.97 on Cora and DBLP, 0.90 on Contact-high-school, and 0.96 on Contact-primary-school. The lowest observed pairwise value is 0.78 on Contact-high-school, indicating greater seed sensitivity on that dataset than on the other three.

D.4 Intervention Controls and Efficiency

Intervention edge ranking. For node v and target type i, the candidate set contains every incident hyperedge that participates in at least one hm-i instance containing v. Candidate scores $s_i^{\mathrm{struct}}(e)$ are computed once on the pre-edit hypergraph and sorted in ascending order. Exact ties are resolved by hyperedge identifier. Hyperedges are deleted individually, rather than as complete motif-instance groups, until $\lceil r N_{v,i} \rceil$ of the $N_{v,i}$ candidates have been removed. MoCHy then recounts all affected h-motif instances and recomputes the normalized structural encodings, while the trained model parameters remain fixed.

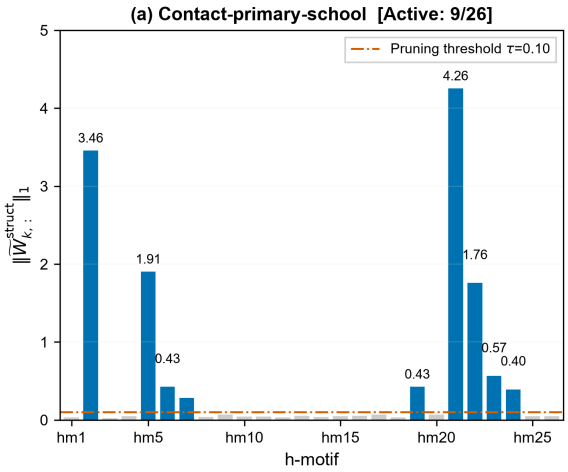


Figure 4: Effective h-motif subset for Contact-primary-school, the counterpart of Figure 2 deferred to the appendix.

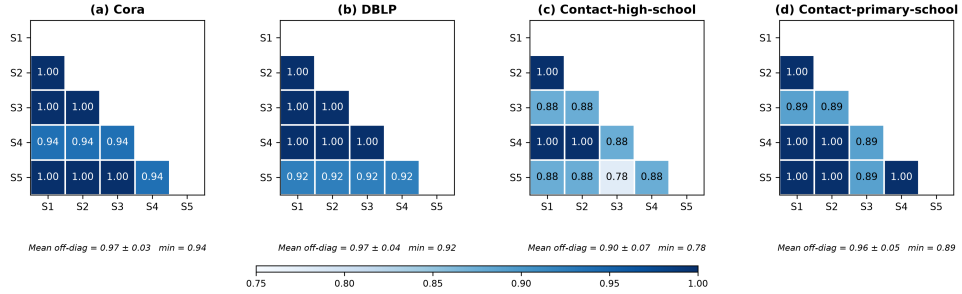


Figure 5: Seed stability of the effective h-motif subset across five training runs: pairwise Jaccard similarity of the selected h-motif sets (minimum 0.78 across datasets).

Null-intervention control. To assess whether the observed prediction changes are specific to the planted signal h-motifs, we additionally intervene on structural components that do not directly enter the predefined ground-truth score. For each evaluated target node, the null intervention uses the same deletion ratio as the corresponding target-signal intervention. All h-motif instances and normalized structural encodings are recomputed after deletion, so a null intervention may still alter hm8 or hm17 indirectly through motif overlap and normalization spillover. It should therefore be interpreted as a specificity control rather than a zero-effect intervention. At $r = 0.5$, the mean absolute logit change under null interventions is approximately one-seventh of that under target-signal interventions (0.10 vs. 0.71). This comparison is descriptive because the current results do not report the paired samples or a signed-rank test.

The closed-form intervention runs in milliseconds and is 36–60$\times$ faster than the counterfactual search of CF-HyperGNNExplainer, while the Feature-Linkage Ratio is at most 0.15. Thus, the structural component accounts for most of the decomposed intervention effect in these two settings. FLR is not an approximation-error metric and, by itself, does not establish agreement with the ground-truth score changes.

D.5 Sensitivity Analysis

This analysis probes how the explanation quality depends on the four main hyperparameters of IV-SENN—the number of feature prototypes K_{feat}, the number of compositional bases K_{comp}, the minimum attention temperature $\tau_{\min}$, and the ℓ_1 sparsity weight λ. We vary one parameter at a time

Table 6: Intervention controls and efficiency on N-SYNTH and Cora. SR_v is the measured node-level spillover of an intervention; FLR is the fraction of the intervention effect attributable to feature linkage.

Dataset	SR_v	FLR	Head time	Speedup vs. CF
N-SYNTH	0.12	0.08	2 ms	60×
Cora	0.25	0.15	5 ms	36×

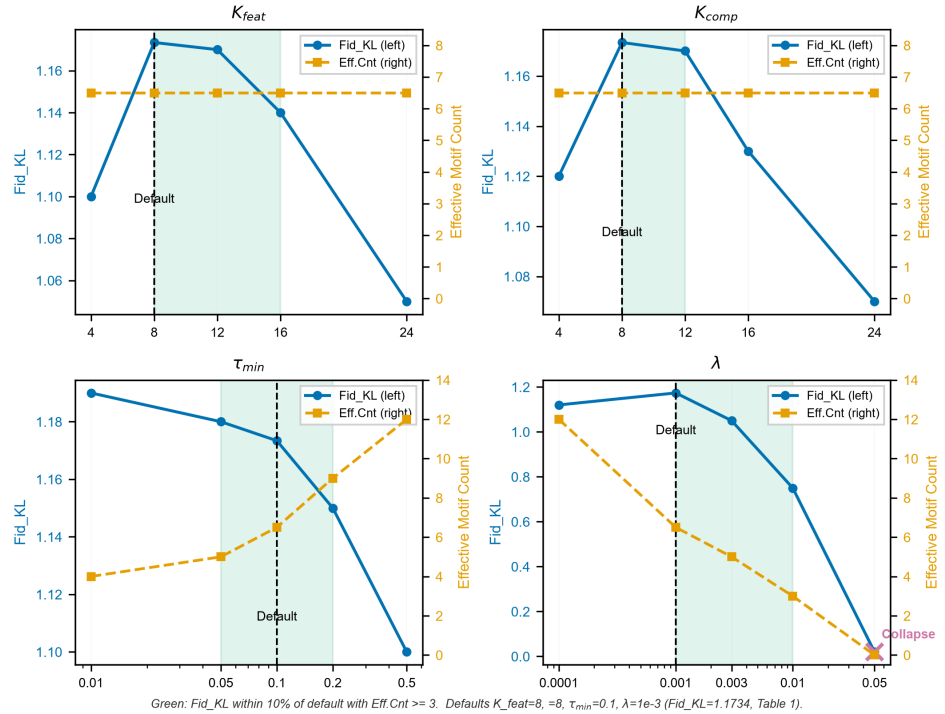


Figure 6: Hyperparameter sensitivity on Cora (five-seed means): one-at-a-time sweeps around the reported configuration ($K_{\text{feat}} = 8$, $K_{\text{comp}} = 8$, $\tau_{\min} = 0.1$, $\lambda = 10^{-3}$; Fid_KL 1.1734, Table 1). Green regions indicate settings within 10% of the default Fid_KL and with effective motif count ≥ 3. At $\lambda = 5 \times 10^{-2}$, excessive ℓ_1 regularization drives the weights to zero (near-uniform prediction).

around the reported configuration ($K_{\text{feat}} = 8$, $K_{\text{comp}} = 8$, $\tau_{\min} = 0.1$, $\lambda = 10^{-3}$), at which Fid_KL equals the Cora value of Table 1 (1.1734) and the effective motif count is 6.5 in the Cora sweep.

For both capacity parameters, Fid_KL peaks at the default value of 8 and declines when the dimension is reduced to 4 or increased to 24. The effective motif count remains at 6.5 throughout these two sweeps. These curves support an intermediate-capacity choice, but they do not by themselves identify underfitting or overfitting as the cause of the decline.

Increasing $\tau_{\min}$ from 0.01 to 0.5 is associated with a decrease in Fid_KL from approximately 1.19 to 1.10 and an increase in the effective motif count from 4 to 12. Since $\tau_{\min}$ controls the lower bound rather than imposing a single temperature, this sweep characterizes the attainable degree of sharpening in explanation extraction. The additive prediction head remains unchanged. The default $\tau_{\min} = 0.1$ lies between the sharpest and most diffuse settings. Because the figure does not include uncertainty intervals for the sweep, small differences between adjacent settings should be interpreted cautiously. The sparsity weight has the largest effect in the displayed range. Weak regularization leaves many components active, whereas progressively stronger regularization reduces both Fid_KL and the effective motif count. At $\lambda = 5 \times 10^{-2}$, both quantities approach zero, consistent with a collapsed, near-uniform predictor. The default $\lambda = 10^{-3}$ attains the highest displayed Fid_KL with an effective motif count of 6.5. This pattern agrees qualitatively with the DBLP ℓ_1-removal ablation

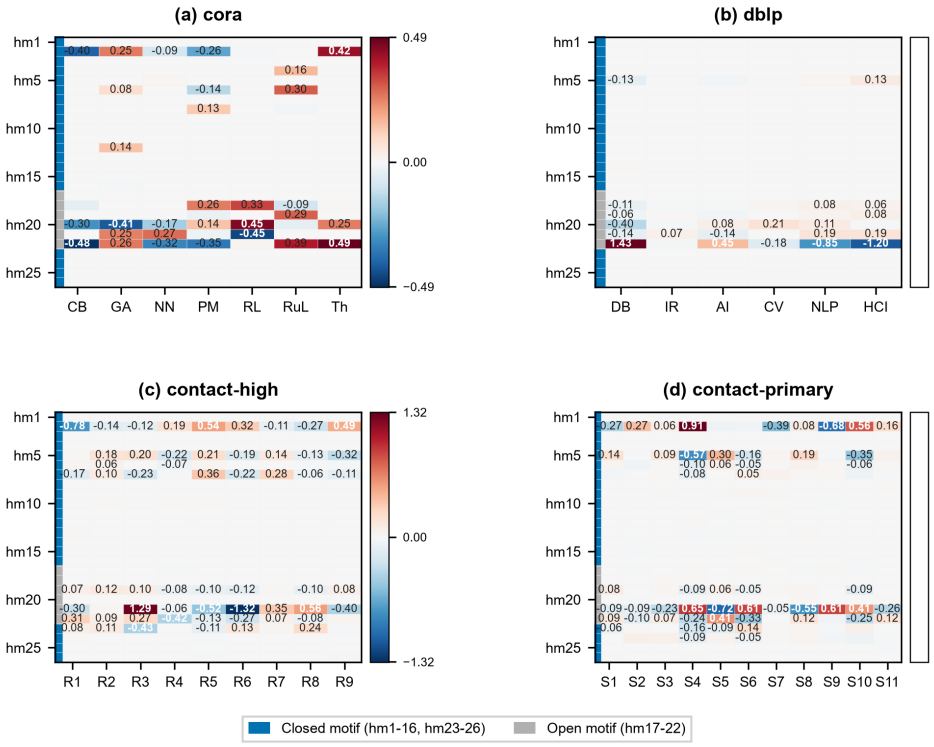


Figure 7: Linear-term component of the global explanation: class-specific h-motif weights read directly from W^{linear}.

Table 7: Local explanation of Cora node 785. Signed ϕ values are predictive contributions, whereas β values are class-agnostic local role evidence and are not added to the class logit.

Quantity	Concept or group	Value
Positive $\phi^{(4)}$	hm8	$+0.842$
	feature prototype 3	$+0.516$
	hm20	$+0.287$
Negative $\phi^{(4)}$	hm17	-0.263
	feature prototype 6	-0.141
Residual contribution	remaining 29 concepts (net)	$+0.639$
Role evidence β	hm8	0.421
	hm20	0.183
	hm21	0.142
	hm17	0.096
	remaining 22 types (total)	0.158

in Table 3 and the N-SYNTH intervention analysis in Section 5.4: weaker sparsity increases the effective motif count and reduces signal h-motif precision, respectively.

D.6 Existence Verification on Real Data

As a concrete recounting check, Cora node 785 participates in an hm8 instance formed by $e_2 = \{783, 785\}$, $e_{777} = \{783, 785, 1513\}$, and $e_{812} = \{785, 1513\}$. Their non-empty Venn regions are AB, BC, and ABC, which match the closed hm8 pattern. This example verifies that the enumeration pipeline recovers a valid h-motif from the original incidence structure. It serves only as a recounting sanity check and does not estimate the prevalence of individual h-motif types.

Table 8: Grounding of the selected hm8 instance for Cora node 785. Node and hyperedge identifiers use the zero-based indexing of the data file.

Edge	Members	$t^{\text{struct}}_{e,\text{hm8}}$	$s^{\text{struct}}_{\text{hm8}}(e)$	$\max_j s^{\text{feat}}_j(e)$
e_2	$\{783, 785\}$	0.780	0.686	0.760
e_{777}	$\{783, 785, 1513\}$	0.920	0.715	0.820
e_{812}	$\{785, 1513\}$	0.710	0.670	0.790

Table 9: Predictive distributions for the local Cora case. "Retain" keeps only the selected h-motif instance; "Remove" deletes its three hyperedges. Encodings are recomputed after each edit.

Graph	p_0	p_1	p_2	p_3	p_4	p_5	p_6	Pred.
Full	0.025	0.030	0.028	0.045	**0.842**	0.018	0.012	4
Retain	0.035	0.042	0.031	0.056	**0.798**	0.023	0.015	4
Remove	0.065	0.073	0.058	**0.418**	0.336	0.029	0.021	3

D.7 Global Explanation Visualization and Local Case Study

Figure 7 visualizes the structural block of the linear term. Most entries are close to zero, while a smaller set of h-motif dimensions has class-dependent positive or negative weights. The active rows and their magnitudes differ across the co-authorship and contact datasets, indicating that the learned linear structural semantics are dataset-specific. Because the figure excludes the compositional and interaction terms, it is a diagnostic of the dominant linear component rather than a complete decomposition of the class logits.

Local explanation of Cora node 785. We extend the recounting example in Appendix D.6, which fixes the node and motif instance independently of the model outputs. Node 785 has ground-truth class 4 (zero-based indexing), and the reported checkpoint also predicts class 4 with probability 0.842. Its predicted-class logit is $z_{785,4} = 2.110$. Table 7 reports the largest signed concept contributions and the dominant local role evidence. The displayed concept contributions, including the aggregate of the unlisted concepts, sum to 1.880; with $b_4 = 0.230$, they satisfy the completeness check $\sum_i \phi^{(4)}_{785,i} + b_4 = 2.110$.

The largest positive structural contribution and the largest role-evidence mass both correspond to hm8. The grounding rule in Eq. 27 selects $\mu_{785} = (e_2, e_{777}, e_{812})$. Table 8 shows the three constituent hyperedges and their matches. The structural matches obey the definition $s^{\text{struct}}_{\text{hm8}}(e) = \sigma(t^{\text{struct}}_{e,\text{hm8}})$ and therefore lie in $[0.5, \sigma(1)]$. The instance match and final grounding score are

$$L_{785}(\mu_{785}) = 0.7402, \tag{46}$$

$$J^{(4)}_{785,\mu_{785}} = 0.842 \times 0.421 \times 0.7402 = 0.2624. \tag{47}$$

The three hyperedges have non-empty regions $AB = \{783\}$, $BC = \{1513\}$, and $ABC = \{785\}$, so the target node occupies the ABC role. Feature prototype 3 has the strongest positive feature contribution, and e_{777} is its highest-matching hyperedge within the selected instance. Thus, the feature prototype annotates an h-motif-grounded explanation rather than selecting an unrelated hyperedge.

Table 9 reports the full, retained, and removed predictive distributions. Retaining only the selected hm8 instance preserves class 4 and changes its probability from 0.842 to 0.798. Removing the same three hyperedges changes the prediction to class 3 and reduces the class-4 probability to 0.336. The corresponding class-4 logits are 2.110, 2.056, and 1.191, respectively.

For this node, the retained-subgraph accuracy indicator is one, $\text{Fid}^-_v = 0$, and $\text{Fid}^+_v = 1$. The distributional changes are $D_{\text{KL}}(p_v \| p^S_v) = 0.0069$ and $D_{\text{KL}}(p_v \| p^{-S}_v) = 0.5870$. Hence the selected motif instance is sufficient to preserve the prediction in this case, while its removal is decision-changing. The signed contributions explain why the class is supported, $\boldsymbol{\beta}_v$ identifies the locally expressed h-motif role, and the instance-grounding rule identifies where that role occurs. This single-node analysis is illustrative; aggregate explanation quality is evaluated by GSACC and the fidelity metrics.