HYPERGRAPH PATTERN MACHINE: COMPOSITIONAL TOKENIZATION FOR HIGHER-ORDER INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hypergraphs model higher-order relations that drive real-world decisions, from drug prescriptions to recommendations. A central structural signal in such data, beyond what pairwise relations can express, is *interaction compositionality*: whether a higher-order relation is compositional, emergent, or inhibitory with respect to its subsets, i.e., whether a set and its one-element-smaller subset are both observed, only the larger one is observed, or only the smaller one is observed. These patterns are read directly from the observed data. In adverse-event reports, for example, they distinguish drug sets whose reported association persists when a drug is added, appears only for the full set, or is absent after adding a drug; these are reporting patterns, not pharmacological mechanisms. However, existing hypergraph learning methods, which propagate messages over observed hyperedges, do not represent unobserved subsets explicitly and leave this signal implicit. To this end, we propose the Hypergraph Pattern Machine (HGPM), which makes the compositional pattern of subsets an explicit part of the input. It tokenizes observed and unobserved subsets around each target, organizes them in an inclusion DAG, and pretrains an inclusion-aware Transformer with masked reconstruction. On ten hypergraph benchmarks, HGPM matches or exceeds state-of-the-art methods. A qualitative case study on adverse-event reports further illustrates how HGPM separates candidate drugs that are close in feature space but have different recorded patterns.

1 INTRODUCTION

Hypergraphs (Feng et al., 2019; Chien et al., 2022; Wang et al., 2023a; Tang et al., 2024; Gao et al., 2026; Ma et al., 2025a) model higher-order relations, e.g., drug interactions, recommendation, knowledge bases. Hypergraph data carries structure that pairwise graphs (Kipf & Welling, 2017; Veličković et al., 2018) cannot express (Hajij et al., 2023; Bodnar et al., 2021b;a). One such property, read directly from the observed data, is whether a higher-order interaction is *compositional*, *emergent*, or *inhibitory* with respect to its observed or absent subsets (Figure 1). We call this property *interaction compositionality*. Yet existing methods (Yadati et al., 2019; Bai et al., 2021; Dong et al., 2020), which propagate messages only over observed hyperedges, do not represent absent subsets explicitly. We argue that this pattern should be an explicit part of the representation, and we expose it to the model as input structure computed from the observed hyperedges alone.

Polypharmacy (Tatonetti et al., 2012; Chou, 2006) offers an intuition. Consider adverse-event reports, where each hyperedge is a set of drugs co-reported with an adverse event. If both a pair and the full triple of drugs are reported, the association persists when the third drug is added (compositional). If only the full triple is reported while the pair is not, the association appears only jointly (emergent). If a pair is reported but the triple is not, the association is absent once the third drug is added (inhibitory). These are data-level reporting patterns, not evidence of pharmacological synergy or suppression, which experimental studies quantify from measured drug responses (Tekin et al., 2018; Lozano-Huntelman et al., 2021).

Existing hypergraph methods, however, do not represent these three patterns explicitly. Hyperedge-centric methods such as AllSet (Chien et al., 2022) and ED-HNN (Wang et al., 2023a) propagate between nodes and observed hyperedges; expansion-based methods such as HGNN (Feng et al., 2019) and LEGCN (Yang et al., 2022) reduce the hypergraph to a clique- or line-expanded pairwise graph and propagate over its edges; simplicial-complex networks (Ebli et al., 2020; Bodnar

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

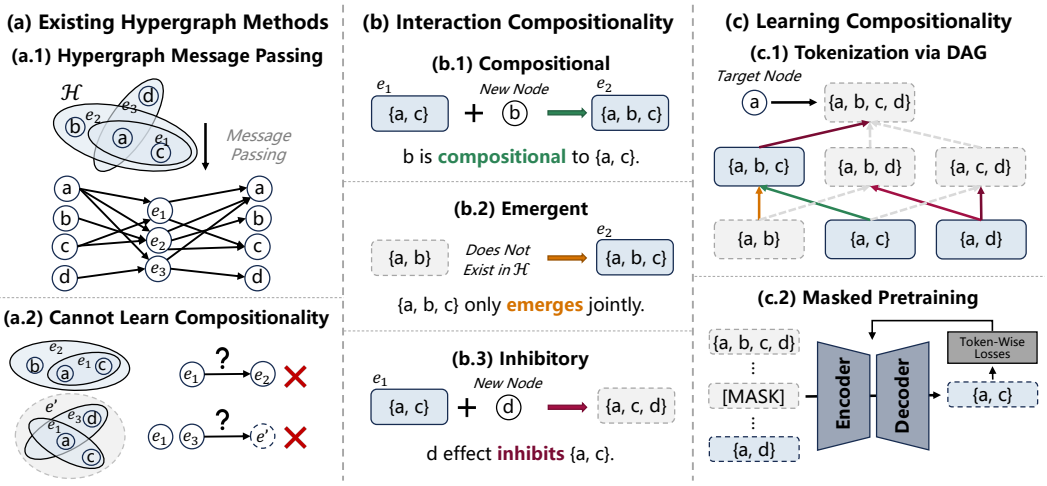


Figure 1: **Interaction compositionality.** **(a)** Existing hypergraph methods propagate messages only over observed hyperedges (a.1), failing to distinguish hyperedge pairs whose compositionality structure differs (a.2). **(b)** Each adjacent-order subset pair falls into one of three regimes by which endpoints are observed: *compositional* (b.1), *emergent* (b.2), and *inhibitory* (b.3). **(c)** HGPM tokenizes observed (solid) and unobserved (dashed) subsets containing a target node into an inclusion DAG with COMP/EMER/INHIB-labeled transitions (c.1, color-matched to (b)); the encoder is pretrained via masked-subset reconstruction (c.2).

et al., 2021b) pass messages between adjacent observed k-faces. These methods rely on message passing (Telyatnikov et al., 2025), where information flows only between entities in the hyperedges that already exist. Yet emergent and inhibitory regimes are defined by which subsets do *not* exist, a signal that is implicit in the observed structure but never represented as a state by these models. With uninformative node features, incidence-based message passing provably cannot separate some hypergraphs whose patterns differ (Appendix B.2).

We therefore make these patterns explicit in the representation. Capturing them raises three challenges. (1) Since absence is itself part of the compositional signal, both observed and absent subsets must become first-class entities the model can read. We tokenize subsets, each token carrying its content, order, and an explicit existence label that distinguishes observed from unobserved. (2) Each subset participates in many supersets, and each adjacent subset–superset transition carries its own label, so the representation must encode all of these multi-parent relations together. We organize tokens into an inclusion DAG whose subset–superset edges carry these labels as first-class structure. (3) Subsets grow combinatorially with the size of the higher-order interaction, making the full DAG intractable on real datasets. We sample tokens within a bounded per-order budget, keeping the DAG local around each target and retaining the subset–superset edges among the sampled tokens. Together, these three choices yield a representation in which every sampled adjacent transition exposes the joint pattern of presence and absence that message passing leaves invisible.

We propose the Hypergraph Pattern Machine (HGPM), a Transformer that learns from interaction compositionality on a target-centered inclusion DAG. For each target entity, HGPM tokenizes a bounded sample of observed and unobserved subsets that contain it; each token carries an existence label, and each subset–superset edge with an observed endpoint carries one of three composition labels: compositional, emergent, or inhibitory. The encoder augments self-attention with structural biases (Ying et al., 2021; Kreuzer et al., 2021) that inject inclusion topology (a direction bias) and the existence roles of token pairs (a role-pair bias), which together identify the regime of each transition on DAG edges. We pretrain HGPM by masking subset tokens (Hou et al., 2022; Xia et al., 2023; Hou et al., 2023) and regressing their representations onto those of an exponential-moving-average (EMA) teacher (Grill et al., 2020; Baevski et al., 2022), optionally with an auxiliary loss that predicts whether each subset is observed. On standard hypergraph node classification benchmarks, HGPM matches or exceeds state-of-the-art methods. Notably, on the HODDI (Wang et al., 2025c) and JADER (Tsuchiya et al., 2020) drug interaction datasets, HGPM outperforms the seven baselines under both edge classification and link prediction. A qualitative case study shows how HGPM separates two add-on drugs that are close in feature space but have opposite recorded patterns.

2 Related Work

Message-Passing-Based Hypergraph Learning. As in graph learning (Gilmer et al., 2017; Kipf & Welling, 2017), most hypergraph methods rely on message passing over observed hyperedges. *Hyperedge-centric methods* (Yadati et al., 2019; Dong et al., 2020; Huang & Yang, 2021; Chien et al., 2022; Wang et al., 2023a; Ma et al., 2025b; 2026a) aggregate between nodes and hyperedges, with recent variants using sheaf, energy, framelet, heat-flow, or higher-order propagation (Duta et al., 2023; Wang et al., 2023b; Li et al., 2025; Xie et al., 2025; Sun et al., 2026). *Expansion-based methods* (Yang et al., 2022; Feng et al., 2019; Ma et al., 2025a) run standard message passing (Veličković et al., 2018; Kipf & Welling, 2017) on a pairwise expansion, while *simplicial-complex networks* (Bodnar et al., 2021b; Eijkelboom et al., 2023) propagate over k-faces but require subset closure, foreclosing emergent transitions. Since all of these pass messages over observed structure (Telyatnikov et al., 2025), the joint pattern of subset presence and absence is never represented; HGPM instead tokenizes both observed and unobserved subsets, making that pattern an explicit input.

Substructure Tokenization for Structured Data. A parallel line in graph learning replaces message passing with substructure tokens encoded by Transformers with structural attention biases (Ying et al., 2021; Rampášek et al., 2022) and masked-reconstruction pretraining (Hou et al., 2022; Xia et al., 2023); this has proved a scalable alternative on graphs (Wang et al., 2024; 2025a;b) and temporal graphs (Ma et al., 2026b). HGPM extends it to hypergraphs: tokens shift from subgraph patterns to centered subset hierarchies that explicitly include unobserved subsets. Hypergraph foundation models (Gao et al., 2026) instead pretrain vertex-centric encoders for cross-corpus transfer. Self-supervised hypergraph methods contrast augmented views (Lee & Shin, 2023) or fill in missing members of hyperedges (Kim et al., 2024), and hyperedge prediction scores candidate sets against sampled negatives, whose choice strongly affects evaluation (Patil et al., 2020; Hwang et al., 2022); simplicial closure (Benson et al., 2018) studies how subsets become jointly observed. In HGPM, unobserved subsets are input tokens rather than training negatives.

Higher-Order Drug Interaction Prediction. Drug-interaction modeling has mostly been pairwise: synergy estimation (Preuer et al., 2018; Yang et al., 2021), side-effect classification (Zitnik et al., 2018; Ryu et al., 2018) including latent-combination models (Nguyen et al., 2022), and substructure-based DDI (Nyamabo et al., 2021; Chen et al., 2021). A more recent line extends to combinations of arbitrary size via permutation-invariant pooling (Peng & Ning, 2019), or hypergraph networks reduced to pairwise targets (Saifuddin et al., 2023), with HODDI (Wang et al., 2025c) and JADER (Tsuchiya et al., 2020) supplying higher-order drug-effect data at scale. Most of these methods score a combination as a single observation, without contrasting it with the indicators of its subsets and adjacent supersets, which HGPM exposes.

3 Problem Setup

We consider a hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, where $\mathcal{V}$ is a set of n entities, $\mathcal{E} \subseteq 2^{\mathcal{V}}$ is the set of observed hyperedges visible to the learner, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ collects per-entity features. For transductive node classification all hyperedges are visible and only node labels are split; for the link tasks only training records are visible. For any subset $S \subseteq \mathcal{V}$, let $\mathbb{1}_S = \mathbb{1}[\, S \in \mathcal{E}\,] \in \{0, 1\}$ denote its observation indicator, with $\mathbb{1}_S = 1$ when S is an observed hyperedge and 0 otherwise. Standard hypergraph methods learn an encoder f_θ from downstream labels such as the class of a node $c \in \mathcal{V}$ or the effect of a combination $G \subseteq \mathcal{V}$; we additionally expose to f_θ the *interaction compositionality* structure of $\mathcal{H}$, defined next, which is computed from $\mathcal{E}$ alone and uses no external or domain knowledge.

Interaction Compositionality. The *inclusion hierarchy* of $\mathcal{H}$ is the set of subsets of $\mathcal{V}$ ordered by inclusion, $(2^{\mathcal{V}}, \subseteq)$. Within it, an *adjacent-order pair* is any (S, S') with $S \subset S'$ and $|S'| = |S| + 1$. Each adjacent-order pair has a *composition label* $\mathrm{comp}(S, S')$ determined by the indicators of S and S': it is Comp if both are observed ($\mathbb{1}_S = 1$, $\mathbb{1}_{S'} = 1$), Emer if only the superset is observed ($\mathbb{1}_S = 0$, $\mathbb{1}_{S'} = 1$), and Inhib if only the subset is observed ($\mathbb{1}_S = 1$, $\mathbb{1}_{S'} = 0$); pairs where both S and S' are absent are excluded as uninformative. Collecting all labeled pairs gives the *compositionality structure* $\mathcal{C}(\mathcal{H})$ of $\mathcal{H}$. Labels belong to pairs, not to sets: a set can be compositional with respect to one subset and emergent with respect to another. Because every unobserved subset counts as absent, labels describe what is recorded, not what is possible. For example, in Figure 1(c.1), with target a and observed hyperedges $\{a, c\}$, $\{a, b, c\}$, and $\{a, d\}$, the pair $(\{a, c\}, \{a, b, c\})$ is

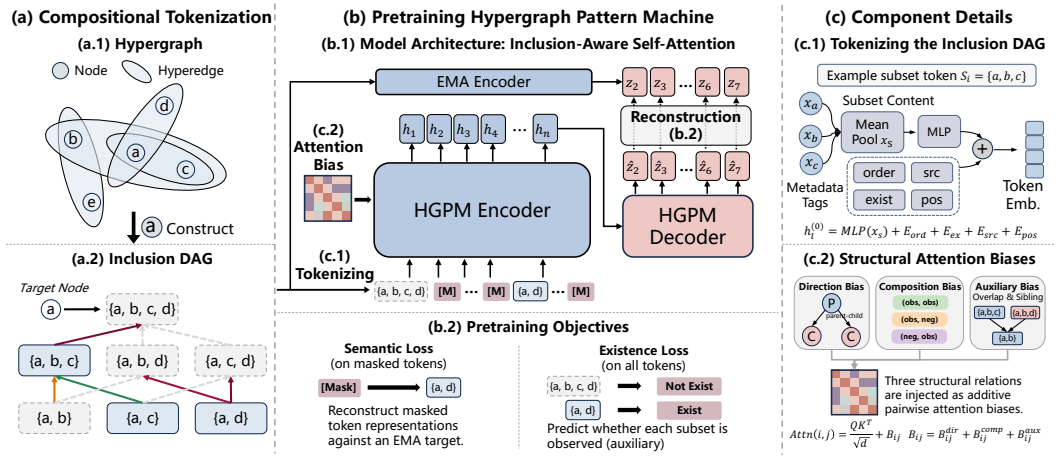


Figure 2: **Hypergraph Pattern Machine.** For each target entity, HGPM tokenizes subsets containing it and organizes them as an inclusion DAG with composition-labeled edges (a). A masked-reconstruction objective supervises subset semantics, optionally with auxiliary existence prediction (b). Pairwise structural biases inject inclusion topology and existence roles into self-attention (c).

COMP, $(\{a, b\}, \{a, b, c\})$ is EMER, and $(\{a, d\}, \{a, b, d\})$ is INHIB. HGPM exposes this structure to the encoder as input rather than predicting it.

Why This Requires Going Beyond Message Passing. Recovering $\mathrm{comp}(S, S')$ requires reading the indicator pair $(\mathbb{1}_S, \mathbb{1}_{S'})$ jointly. This is not guaranteed for hyperedge message-passing encoders bounded by the incidence Weisfeiler–Leman test, whose representations propagate only through observed hyperedges: with uninformative node features, there exist hypergraph pairs on which any such encoder produces identical features at every finite depth (Xu et al., 2019; Morris et al., 2019), yet whose compositionality structures differ (Appendix B.2). What is needed instead is a representation that treats every relevant subset, observed or not, as a first-class object, with its existence indicator readable directly from the input. Under an idealized tokenizer, such a representation separates, for almost all weights, a non-isomorphic pair on which all such encoders coincide (Appendix B.3).

4 HYPERGRAPH PATTERN MACHINE

We propose HGPM, an inclusion-aware Transformer that learns from a target-centered inclusion DAG of subset tokens (Figure 2). Concretely, for each target $c \in \mathcal{V}$ HGPM proceeds in three stages: (i) build a bounded inclusion DAG $\mathcal{G}_c$ of subset tokens around c, with composition labels read directly from observation indicators (Section 4.1); (ii) linearize $\mathcal{G}_c$ and encode it with self-attention augmented by pairwise biases that re-inject the inclusion topology and existence roles at every layer (Section 4.2); (iii) pretrain the encoder with masked reconstruction, then transfer it via lightweight readouts (Section 4.3).

4.1 COMPOSITIONAL TOKENIZATION

Inclusion DAG. For each target entity $c \in \mathcal{V}$, HGPM organizes the local inclusion hierarchy around c as a directed acyclic graph (DAG) $\mathcal{G}_c = (\mathcal{S}_c, \mathcal{R}_c)$ with existence-labeled nodes and composition-labeled edges, exposing a target-centered part of the compositionality structure $\mathcal{C}(\mathcal{H})$ in the input:

$$\begin{aligned} \mathcal{S}_c &= \{\, S \subseteq \mathcal{V} : c \in S,\ 1 \le |S| \le K_{\max} \,\}, \\ \mathcal{R}_c &= \{\, (S, S') \in \mathcal{S}_c \times \mathcal{S}_c : S \subset S',\ |S'| = |S| + 1 \,\}. \end{aligned} \tag{1}$$

The node set contains only subsets that include c, anchoring every token and edge to the prediction target. The edge set keeps only adjacent-order pairs, since compositionality is defined locally on (S, S') with $|S'| = |S| + 1$ and non-adjacent inclusions arise as chains of atomic steps. Each edge with an observed endpoint carries the composition label $\mathrm{comp}(S, S')$, read off the data without learning, which promotes pair-level compositionality from a derived quantity to a first-class structural object. We use a DAG rather than a tree because a subset typically has multiple adjacent-order supersets carrying different labels. For example, if a corpus records $\{\mathrm{aspirin}, \mathrm{warfarin}\}$ and

$\{\text{aspirin}, \text{warfarin}, \text{ibuprofen}\}$ but not $\{\text{aspirin}, \text{warfarin}, \text{vitamin K}\}$, the pair extends to the first triple as COMP and to the second as INHIB, and a tree rooted at the pair must commit to one child label and erase the other. In practice, $|\mathcal{S}_c|$ grows combinatorially in $|\mathcal{V}|$. We therefore cap how many tokens we keep at each order, filling the cap with observed subsets first and unobserved ones second, so the tokenized DAG is a sampled subset of $\mathcal{S}_c$ (Appendix C.1).

Tokenizing the Inclusion DAG. To feed $\mathcal{G}_c$ to a Transformer, we first linearize it into a unique sequence by sorting subsets in descending $(|S|, S)$ order, placing the center singleton $\{c\}$ at the last position (we draw V such samples, called views, and concatenate them, each ending with its own center token); the edge set $\mathcal{R}_c$ is dropped here and re-injected as structural attention biases in Section 4.2. Each position in the sequence is a token for one subset $S \in \mathcal{S}_c$, carrying four attributes:

$$\mathbf{t}_S = \big(\mathbf{x}_S,\, |S|,\, \mathbb{1}_S,\, \tau(S)\big), \quad \mathbf{x}_S = \frac{1}{|S|}\sum_{v \in S} \mathbf{x}_v, \quad \tau(S) \in \{\text{center}, \text{obs}, \text{sub}, \text{neg}\}. \tag{2}$$

Each attribute is chosen to expose a distinct signal that compositionality depends on. The *semantic* feature $\mathbf{x}_S$ is mean-pooled for permutation-invariance (Zaheer et al., 2017). The *order* $|S|$ lets the model condition on hierarchy depth. The *existence indicator* $\mathbb{1}_S$ is kept first-class because compositionality is defined by existence patterns alone, and so we can mask it during pretraining (Section 4.3). The *source* $\tau(S)$ tags each token as center, obs, sub, or neg. An obs token is an observed hyperedge containing c with at most $K_{\max}$ members. A sub token is a center-anchored subset of an observed hyperedge, sampled as a random nested chain ($\{c\}$ plus a growing prefix of the remaining members); sub tokens are how hyperedges with more than $K_{\max}$ members enter the DAG, and their existence input is set to 1 (contained in an observed hyperedge). A neg token is an unobserved subset, exposing absence as a signal. They are needed because EMER and INHIB transitions each have an absent endpoint. We construct them by *local perturbation* of an observed hyperedge or sub token e containing c: drop a non-center member, add a node from $\mathcal{V} \setminus e$, or replace a non-center member with one from $\mathcal{V} \setminus e$; retain only if order $\in [2, K_{\max}]$ and S is neither a visible hyperedge nor a sub token of the same view. Neg tokens are inputs describing the neighborhood of c; they carry no task label and are distinct from the negative records of the drug tasks.

4.2 INCLUSION-AWARE SELF-ATTENTION

The encoder is a Transformer over the linearized token sequence, with each token S entering as $\mathbf{h}_S^{(0)} = W_{\text{x}}(\mathbf{x}_S) + E_{\text{ord}}[|S|] + E_{\text{ex}}[\mathbb{1}_S] + E_{\tau}[\tau(S)] + E_{\text{view}}[v(S)] + E_{\text{pos}}[\pi(S)]$, a feature-MLP projection summed with learned lookups E for the four token attributes, its view $v(S)$, and its linearized position $\pi(S)$ (Ying et al., 2021; Rampášek et al., 2022); the stacked embeddings form $\mathbf{H}_c^{(0)} \in \mathbb{R}^{|\mathcal{S}_c| \times d}$. Linearization, however, discards $\mathcal{R}_c$, so we re-inject inclusion topology and existence roles by augmenting self-attention with pairwise structural biases at every layer.

Structural Attention Biases. At layer $\ell \in \{1, \ldots, L\}$, each self-attention layer augments the standard attention logit between tokens i and j with a learned pairwise bias (Ying et al., 2021; Kreuzer et al., 2021),

$$a_{ij} = \frac{\mathbf{q}_i^{\top}\mathbf{k}_j}{\sqrt{d_h}} + b_{ij}^{\text{dir}} + b_{ij}^{\text{comp}} + b_{ij}^{\text{aux}}, \tag{3}$$

where $\mathbf{q}_i, \mathbf{k}_j \in \mathbb{R}^{d_h}$ are the per-head query/key projections of the layer-ℓ token states $\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)}$ (d_h is the head dimension). The *direction* bias b_{ij}^{dir} encodes the inclusion edge between S_i and S_j, taking one of three values (no edge, S_i is a parent of S_j, or S_j is a parent of S_i), making the topology of $\mathcal{R}_c$ visible to attention. The *composition* bias b_{ij}^{comp} is indexed by the ordered joint existence roles $(\tau(S_i), \tau(S_j))$ of the query and key tokens and is shared by all token pairs with the same roles, adjacent or not. It does not identify the composition label by itself: the (obs, neg) bucket, for example, covers an EMER edge when the query is the superset and an INHIB edge when it is the subset, so the label of an adjacent pair is determined by b^{comp} together with b^{dir}. The auxiliary term b_{ij}^{aux} folds in non-load-bearing structural priors (order, overlap, sibling) detailed in Appendix C.2. All biases are learned, head-specific, and added at every layer (shared across layers), yielding contextualized representations $\mathbf{H}_c^{(L)} \in \mathbb{R}^{|\mathcal{S}_c| \times d}$ after L stacked layers.

Table 1: **Hypergraph node classification.** Test accuracy on eight benchmarks, split into homophilic (left) and heterophilic (right). Baseline results are taken from published papers (Chien et al., 2022; Wang et al., 2023a; Duta et al., 2023; Wang et al., 2023b; Li et al., 2025; Xie et al., 2025; Sun et al., 2026). AR denotes each method's mean rank across the datasets it reports (lower is better). Best / second / third per column.

	Homophilic				Heterophilic				AR↓
Method	Citeseer	Pubmed	Cora-CA	DBLP-CA	Congress	Senate	Walmart	House	
MLP (Chien et al., 2022)	$72.7_{\pm 1.6}$	$87.5_{\pm 0.5}$	$74.3_{\pm 1.9}$	$84.8_{\pm 0.2}$	–	–	$45.5_{\pm 0.2}$	$67.9_{\pm 2.3}$	12.5
CEGAT (Chien et al., 2022)	$70.6_{\pm 1.3}$	$86.8_{\pm 0.4}$	$76.2_{\pm 1.2}$	$88.6_{\pm 0.3}$	–	–	$51.1_{\pm 0.6}$	$69.1_{\pm 3.0}$	12.8
HGNN (Feng et al., 2019)	$72.4_{\pm 1.2}$	$86.4_{\pm 0.4}$	$82.6_{\pm 1.7}$	$91.0_{\pm 0.2}$	$91.3_{\pm 1.2}$	$48.6_{\pm 4.5}$	$62.0_{\pm 0.2}$	$61.4_{\pm 3.0}$	11.6
HyperGCN (Yadati et al., 2019)	$71.3_{\pm 0.8}$	$82.8_{\pm 8.7}$	$79.5_{\pm 2.1}$	$89.4_{\pm 0.3}$	$55.1_{\pm 2.0}$	$42.5_{\pm 3.7}$	$44.7_{\pm 2.8}$	$48.3_{\pm 2.9}$	13.6
HNHN (Dong et al., 2020)	$72.6_{\pm 1.6}$	$86.9_{\pm 0.3}$	$77.2_{\pm 1.5}$	$86.8_{\pm 0.3}$	$53.4_{\pm 1.5}$	$50.9_{\pm 6.3}$	$47.2_{\pm 0.4}$	$67.8_{\pm 2.6}$	12.4
UniGCNII (Huang & Yang, 2021)	$73.1_{\pm 2.2}$	$88.3_{\pm 0.4}$	$83.6_{\pm 1.1}$	$91.7_{\pm 0.2}$	**$94.8_{\pm 0.8}$**	$49.3_{\pm 4.3}$	$54.5_{\pm 0.4}$	$67.3_{\pm 2.6}$	8.4
AllDeepSets (Chien et al., 2022)	$70.8_{\pm 1.6}$	$88.8_{\pm 0.3}$	$82.0_{\pm 1.5}$	$91.3_{\pm 0.3}$	$91.8_{\pm 1.5}$	$48.2_{\pm 5.7}$	$64.6_{\pm 0.3}$	$67.8_{\pm 2.4}$	9.8
AllSetTransformer (Chien et al., 2022)	$73.1_{\pm 1.2}$	$88.7_{\pm 0.4}$	$83.6_{\pm 1.5}$	$91.5_{\pm 0.2}$	$92.2_{\pm 1.1}$	$51.8_{\pm 5.2}$	$65.5_{\pm 0.3}$	$69.3_{\pm 2.2}$	7.3
ED-HNN (Wang et al., 2023a)	$73.7_{\pm 1.4}$	**$89.0_{\pm 0.5}$**	$84.0_{\pm 1.6}$	$91.9_{\pm 0.2}$	**$95.0_{\pm 1.0}$**	$64.8_{\pm 5.1}$	$66.9_{\pm 0.4}$	$72.5_{\pm 2.3}$	**4.9**
SheafHyperGNN (Duta et al., 2023)	$74.7_{\pm 1.2}$	$87.7_{\pm 0.6}$	**$85.5_{\pm 1.3}$**	$91.6_{\pm 0.2}$	$91.8_{\pm 1.6}$	$68.7_{\pm 4.7}$	–	$73.6_{\pm 2.3}$	6.0
PhenomNN (Wang et al., 2023b)	**$75.1_{\pm 1.6}$**	$88.1_{\pm 0.5}$	**$85.8_{\pm 0.9}$**	**$91.9_{\pm 0.2}$**	$88.2_{\pm 1.5}$	$67.7_{\pm 5.2}$	$63.4_{\pm 0.4}$	$70.7_{\pm 2.4}$	6.0
FrameHGNN (Li et al., 2025)	$74.7_{\pm 2.1}$	$88.7_{\pm 0.4}$	**$85.2_{\pm 0.7}$**	–	–	$67.6_{\pm 5.3}$	–	$72.8_{\pm 2.2}$	5.2
KHGNN (Xie et al., 2025)	$74.8_{\pm 1.1}$	$88.5_{\pm 0.5}$	–	$91.2_{\pm 0.4}$	$92.6_{\pm 1.0}$	$71.2_{\pm 3.9}$	$65.2_{\pm 0.8}$	$74.3_{\pm 2.9}$	5.6
HealHGNN (Sun et al., 2026)	**$75.1_{\pm 1.2}$**	**$88.8_{\pm 0.3}$**	–	**$92.0_{\pm 0.3}$**	$93.5_{\pm 0.8}$	**$76.1_{\pm 4.1}$**	**$68.2_{\pm 0.6}$**	**$77.2_{\pm 2.3}$**	**2.9**
HGPM (no pretrain)	$74.1_{\pm 1.1}$	$87.5_{\pm 0.5}$	$82.3_{\pm 1.7}$	$91.1_{\pm 0.3}$	$92.4_{\pm 0.6}$	**$75.2_{\pm 3.5}$**	**$68.9_{\pm 0.3}$**	**$77.8_{\pm 1.2}$**	6.4
HGPM	**$77.1_{\pm 1.2}$**	**$89.7_{\pm 0.5}$**	$85.1_{\pm 0.2}$	**$92.2_{\pm 0.2}$**	**$94.1_{\pm 1.0}$**	**$77.2_{\pm 4.8}$**	**$71.2_{\pm 0.5}$**	**$79.2_{\pm 1.8}$**	**1.6**

4.3 Compositionality-Aware Pretraining

We pretrain the encoder with a masked reconstruction objective (Hou et al., 2022; Xia et al., 2023; Hou et al., 2023). For each inclusion DAG $\mathcal{G}_c$, we randomly mask a fraction ρ of non-center tokens $\mathcal{M} \subset \mathcal{S}_c \setminus \{\{c\}\}$ by replacing their input embedding with a learnable vector $\mathbf{e}_{\text{mask}}$ (the attention biases are still computed from the token attributes) and forward the corrupted sequence through the encoder, yielding contextualized representations $\tilde{\mathbf{h}}_S^{(L)}$. A prediction head maps $\tilde{\mathbf{h}}_S^{(L)}$ to $\hat{\mathbf{z}}_S \in \mathbb{R}^d$, and the target $\mathbf{z}_S$ is the output, at token S, of a teacher, i.e., an exponential-moving-average (EMA) copy of the encoder and prediction head applied to the *unmasked* sequence (Grill et al., 2020; Caron et al., 2021; Baevski et al., 2022). Two auxiliary heads predict each token's existence indicator ($\hat{\mathbb{1}}_S$) and source tag ($\hat{\tau}_S$). The objective is

$$\mathcal{L} = \underbrace{\frac{1}{|\mathcal{M}|\,d}\sum_{S\in\mathcal{M}} \|\hat{\mathbf{z}}_S - \mathrm{sg}(\mathbf{z}_S)\|_2^2}_{\mathcal{L}_{\text{sem}}} + \lambda_{\text{exist}} \underbrace{\frac{1}{|\mathcal{S}_c|}\sum_{S\in\mathcal{S}_c} \mathrm{BCE}(\hat{\mathbb{1}}_S, \mathbb{1}_S)}_{\mathcal{L}_{\text{exist}}} + \lambda_{\text{src}} \underbrace{\frac{1}{|\mathcal{S}_c|}\sum_{S\in\mathcal{S}_c} \mathrm{CE}(\hat{\tau}_S, \tau(S))}_{\mathcal{L}_{\text{src}}} \tag{4}$$

where sg denotes stop-gradient, $\mathrm{BCE}(p, y) = -y \log p - (1-y)\log(1-p)$ is the binary cross-entropy, and CE is the cross-entropy. Gradients update the student (feature projection, embeddings, bias tables, Transformer blocks), the three heads, and $\mathbf{e}_{\text{mask}}$; the teacher receives no gradient and is updated after every step as $\bar{\theta} \leftarrow m\bar{\theta} + (1-m)\theta$ with $m = 0.996$. On the node-classification benchmarks we use $\lambda_{\text{exist}} = \lambda_{\text{src}} = 0$, i.e., semantic reconstruction only, except on Congress ($\lambda_{\text{exist}} = \lambda_{\text{src}} = 1$); because existence and source are also token inputs, the auxiliary terms act as consistency terms for unmasked tokens. The pretrained encoder transfers to node-level and edge-level downstream tasks via lightweight MLP readouts and end-to-end finetuning, with full readout definitions and the drug-set instantiation of the DAG in Appendix C.3.

5 Generality on Hypergraph Benchmarks

Experimental setup. We evaluate on eight standard hypergraph node classification benchmarks, following the suite established by Chien et al. (2022) and Wang et al. (2023a): four homophilic (Citeseer, Pubmed, Cora-CA, DBLP-CA) and four heterophilic (Congress, Senate, Walmart, House) benchmarks spanning co-citation, co-authorship, co-purchase, and political-collaboration domains with 50/25/25 splits (detailed in Appendix D). We compare HGPM against a comprehensive set of models: MLP, CEGAT (Chien et al., 2022), HGNN (Feng et al., 2019), HyperGCN (Yadati et al., 2019), HNHN (Dong et al., 2020), UniGCNII (Huang & Yang, 2021), AllSet (Chien et al., 2022), ED-HNN (Wang et al., 2023a), SheafHyperGNN (Duta et al., 2023), PhenomNN (Wang et al., 2023b), FrameHGNN (Li et al., 2025), KHGNN (Xie et al., 2025), and HealHGNN (Sun et al., 2026). For

Table 2: **Ablation on the eight hypergraph benchmarks**, reported as homophilic and heterophilic group averages. Gray marks the HGPM default; full per-dataset breakdown in Appendix F.

(a) Subset Tokens

Setting	Homo	Hetero
Observed hyperedges	83.43	75.46
Uncentered random	84.78	77.51
No negatives	85.81	79.61
Inclusion DAG	**86.03**	**80.42**

(b) Inclusion DAG Edges

Setting	Homo	Hetero
No edges	83.66	76.66
Random edges	84.58	78.15
Center-only	85.31	79.23
Adjacent inclusion	**86.03**	**80.42**

(c) Token Encoder

Setting	Homo	Hetero
Mean pooling	83.29	76.29
GRU	79.55	72.91
Vanilla Transformer	84.78	77.99
Inclusion-aware	**86.03**	**80.42**

(d) Structural Attention Biases

Setting	Homo	Hetero
No bias	84.78	77.99
Composition only	85.71	79.85
Direction + aux	85.30	79.10
Full bias	**86.03**	**80.42**

(e) Reconstruction Target

Setting	Homo	Hetero
Existence label	83.34	78.15
Raw feature	84.51	79.13
Composition label	85.18	79.77
EMA-teacher target	**86.03**	**80.42**

(f) Token Sequence Order

Setting	Homo	Hetero
Ascending	85.31	79.08
Shuffled within	85.64	79.67
Fully shuffled	85.89	80.16
Descending	**86.03**	**80.42**

HGPM we report both *HGPM (no pretrain)* trained from scratch and the pretrained-then-finetuned variant. Baseline results are those reported in the cited works. HGPM hyperparameters are tuned by per-dataset random search; the details are presented in Appendix E.1. HGPM results are reported as mean test accuracy over ten random seeds. Despite operating on subset-token sequences, HGPM trains efficiently on a single A40 GPU: on smaller benchmarks (e.g., Cora-CA, Citeseer, Senate), pretraining completes within 2 hours and finetuning within 10 minutes; on larger benchmarks (e.g., Walmart), the full pretrain-and-finetune pipeline completes within 5 hours.

HGPM is consistently strong across homophilic and heterophilic benchmarks. In Table 1, HGPM attains the highest mean test accuracy on six of the eight datasets and the best average rank (1.6, versus 2.9 for the strongest baseline, HealHGNN). Its largest margins are on Walmart (+3.0), Citeseer (+2.0), and House (+2.0); on DBLP-CA and Senate the margin lies within one standard deviation, and on Cora-CA and Congress HGPM is not the best. Pretraining is essential to the full result: without it, HGPM's average rank drops from 1.6 to 6.4.

Compositional structure helps when made explicit. Three independent architectural layers, namely token construction (Table 2a), edge selection (Table 2b), and attention bias (Table 2d), converge on the same conclusion: exposing the subset structure in the input helps. Centered subset tokens beat raw hyperedges most decisively on heterophilic data; center-anchored inclusion edges substantially outperform random edges of identical count, so the *content* of the inclusion DAG matters far more than its density; and the composition bias b^{comp} in Eq. (3) alone recovers 70–80% (mean 75%) of the no-bias-to-full-bias gap on every benchmark. Removing all neg tokens costs only 0.22/0.81 points (Table 2a), so most of the gain comes from the centered subset tokens and inclusion-aware attention. Linearization probes (Table 2f) are consistent with this picture: topological signal flows through the attention bias rather than the position embedding.

Inductive bias alignment matters more than model capacity. A *wrong* inductive bias is worse than no token-to-token interaction at all, and the same pattern recurs at two independent layers. In the subset-token encoder (Table 2c), the GRU baseline underperforms even mean pooling on every dataset because it imposes a sequential ordering on a permutation-invariant subset-token set (Zaheer et al., 2017); the inclusion-aware Transformer's advantage over the vanilla Transformer is therefore inclusion-aware refinement, not extra capacity. The edge-selection ablation (Table 2b) shows the same dynamic from the opposite side: random edges matched in count to the adjacent inclusion edges substantially underperform center-anchored ones, so capacity-matched but bias-misaligned wiring leaves most of the signal on the table. For the four primary removals in (a)–(d), the heterophilic benchmarks see roughly 1.6–2× the impact of the homophilic ones.

The pretraining gain comes from the semantic target. Existence-label reconstruction alone is slightly worse than no pretraining on every benchmark (Table 2e), while the EMA-teacher target $\mathcal{L}_{\mathrm{sem}}$ recovers the full pretrain-vs-no-pretrain gap; raw subset features and discrete composition labels lie in between. The ordering existence → raw feature → composition label → EMA teacher holds on every benchmark, although adjacent steps are within seed variance on the small benchmarks. With the target fixed, the structural biases still add 1.25/2.43 points (Table 2d), so the structural and pretraining contributions are complementary.

Table 3: **Drug interaction prediction.** Test performance on HODDI and JADER, each with edge classification and binary link prediction. AR is the mean rank across the eight metric columns (lower is better). Best / second / third per column.

	HODDI				JADER				
	Edge Class.		Link Pred.		Edge Class.		Link Pred.		AR ↓
Method	F1	AUROC	AUROC	AUPRC	F1	AUROC	AUROC	AUPRC	
MLP	$88.3_{\pm 0.1}$	$93.2_{\pm 0.1}$	$96.6_{\pm 0.1}$	$97.3_{\pm 0.1}$	$74.8_{\pm 0.3}$	$82.3_{\pm 0.0}$	$57.9_{\pm 0.1}$	$56.4_{\pm 0.3}$	5.8
GAT (Veličković et al., 2018)	$88.3_{\pm 0.3}$	$93.3_{\pm 0.3}$	$95.3_{\pm 0.1}$	$95.9_{\pm 0.1}$	$72.1_{\pm 0.6}$	$79.3_{\pm 0.1}$	$56.3_{\pm 0.2}$	$55.1_{\pm 0.3}$	7.8
GCN (Kipf & Welling, 2017)	$89.5_{\pm 0.1}$	$94.9_{\pm 0.0}$	$96.1_{\pm 0.1}$	$97.0_{\pm 0.1}$	$73.8_{\pm 0.2}$	$81.1_{\pm 0.2}$	$57.8_{\pm 0.0}$	$56.3_{\pm 0.1}$	6.1
HGNN (Feng et al., 2019)	$89.5_{\pm 0.2}$	$94.9_{\pm 0.1}$	$96.1_{\pm 0.1}$	$97.0_{\pm 0.1}$	$74.3_{\pm 0.4}$	$81.2_{\pm 0.2}$	$57.7_{\pm 0.2}$	$56.2_{\pm 0.1}$	6.1
AllSetTransformer (Chien et al., 2022)	$\mathbf{90.4}_{\pm 0.1}$	$94.9_{\pm 0.0}$	$\mathbf{96.9}_{\pm 0.1}$	$97.6_{\pm 0.1}$	$\mathbf{75.4}_{\pm 0.3}$	$83.4_{\pm 0.1}$	$\mathbf{59.6}_{\pm 0.1}$	$57.5_{\pm 0.1}$	3.6
ED-HNN (Wang et al., 2023a)	$90.2_{\pm 0.1}$	$\mathbf{95.6}_{\pm 0.0}$	$96.9_{\pm 0.1}$	$\mathbf{97.8}_{\pm 0.1}$	$75.3_{\pm 0.2}$	$\mathbf{83.5}_{\pm 0.1}$	$59.1_{\pm 0.1}$	$\mathbf{58.2}_{\pm 0.1}$	**3.3**
KHGNN (Xie et al., 2025)	$\mathbf{90.5}_{\pm 0.1}$	$\mathbf{95.7}_{\pm 0.0}$	$\mathbf{97.2}_{\pm 0.1}$	$\mathbf{97.6}_{\pm 0.1}$	$\mathbf{75.7}_{\pm 0.3}$	$\mathbf{83.7}_{\pm 0.1}$	$\mathbf{59.4}_{\pm 0.1}$	$\mathbf{57.9}_{\pm 0.1}$	**2.4**
HGPM	$\mathbf{92.9}_{\pm 0.1}$	$\mathbf{96.8}_{\pm 0.0}$	$\mathbf{97.9}_{\pm 0.0}$	$\mathbf{98.2}_{\pm 0.0}$	$\mathbf{78.4}_{\pm 0.2}$	$\mathbf{86.2}_{\pm 0.1}$	$\mathbf{63.2}_{\pm 0.1}$	$\mathbf{61.5}_{\pm 0.1}$	**1.0**

6 Compositionality on Drug Interaction

Experimental setup. We evaluate HGPM on two pharmacovigilance corpora whose reporting conventions, drug vocabularies, and label distributions differ substantially; we train and evaluate on each separately. HODDI (Wang et al., 2025c), derived from the U.S. FAERS, covers, after our preprocessing, 1,821 drugs and a 1,710-class side-effect vocabulary with dense per-drug supervision; JADER (Tsuchiya et al., 2020), the Japanese Adverse Drug Event Report database maintained by Japan's PMDA, covers 4,230 drugs over a more concentrated label space of 100 retained classes (per-dataset statistics in Appendix D.2). Each record pairs a drug set G with a side effect s and a binary label, so both tasks below are conditioned on the side effect and scored by the readout in Appendix C.3. Each dataset supports two tasks: *edge classification*, scoring whether an observed drug combination is reported with a given side effect, and *link prediction*, a binary task distinguishing observed combination–side-effect records from negative records. Both use $50/25/25$ random train/validation/test splits. Token existence and composition labels for the drug DAGs are computed from training records only, and the queried set's own label is masked (Appendix C.3), so validation and test records never provide input labels. We compare HGPM against baselines spanning the methodological spectrum currently applied to drug interaction modeling: flat encoders that aggregate regimens as bag-of-drugs (MLP, GAT (Veličković et al., 2018), GCN (Kipf & Welling, 2017)) and hypergraph encoders that treat each regimen as an atomic hyperedge (HGNN (Feng et al., 2019), AllSetTransformer (Chien et al., 2022), ED-HNN (Wang et al., 2023a), KHGNN (Xie et al., 2025)). For edge classification we report F1 (threshold 0.5) and AUROC; for link prediction we report AUROC and AUPRC. All metrics are averaged over ten random seeds.

HGPM ranks first on all eight metrics across both drug benchmarks. HGPM achieves the best result on all eight metric columns of Table 3 (AR $= 1.0$, ahead of KHGNN at 2.4 and ED-HNN at 3.3). On the saturated HODDI benchmark, HGPM advances edge F1 from 90.5 to 92.9; on the harder JADER, whose drug vocabulary more than doubles (4,230 vs. 1,821), link AUROC rises from 59.6 to 63.2. A plausible source of the gains is that HGPM's per-drug inclusion DAGs expose which subsets of the queried set are recorded with the same side effect in the training data, which flat and hyperedge-centric encoders do not represent explicitly; isolating which patterns drive the gains is left to future work.

HGPM generalizes to interaction orders unseen in training. Higher-order drug interaction data is intrinsically heavy-tailed: the combinatorial space grows as $\binom{|\mathcal{V}|}{k}$ and reporting bias toward common regimens leaves the tail sparse. HODDI, for instance, contains roughly 49,000 size-2 records but only 449 size-8 records. The combinations of greatest safety interest are exactly the rare ones (novel polypharmacy regimens, off-label combinations, post-marketing surveillance leads), so cross-order generalization is the deployment scenario itself, not a stress test. Figure 3 reports two regimes on HODDI:

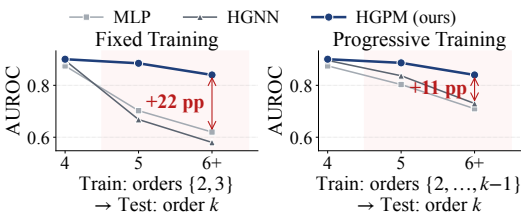


Figure 3: **Cross-order generalization on HODDI.** We report the test AUROC for *Fixed Training* on orders $\{2,3\}$ and *Progressive Training* on $\{2,\ldots,k-1\}$ for each test order k.

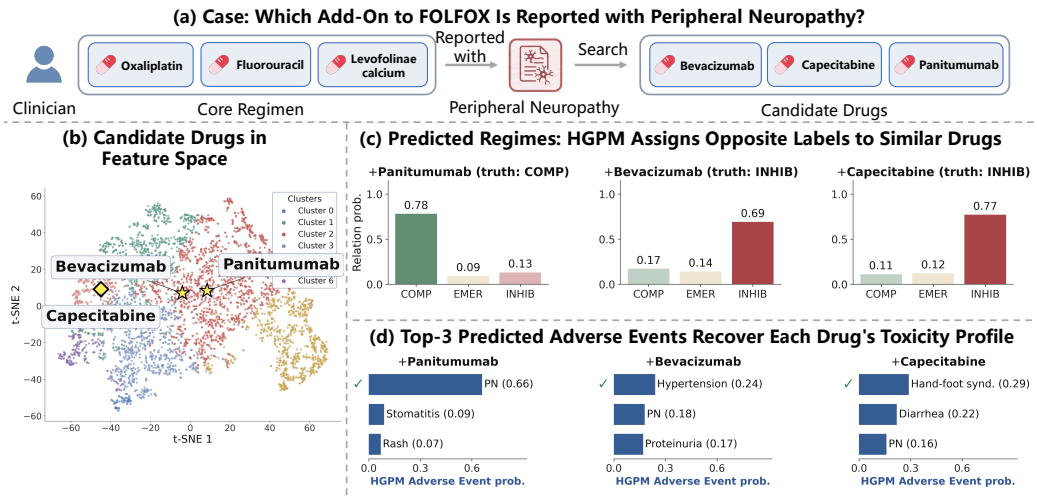


Figure 4: **Case study** on add-ons to FOLFOX with and without a report of peripheral neuropathy.

Fixed Training on orders $\{2, 3\}$, and *Progressive Training* on $\{2, \ldots, k-1\}$ for each test order k. The baselines (MLP and HGNN), whose order-k predictions can only be supervised through actual order-k training edges, collapse under Fixed and trail HGPM by up to 22 pp at order $6+$; under Progressive the gap narrows to 11 pp. HGPM's curve is much flatter across both regimes, because the inclusion DAG routes every high-order target through a chain of inclusion edges to lower-order subsets that are dense in training, transporting the dense low-order signal into the sparse tail.

Case study: HGPM separates feature-similar drugs with opposite recorded outcomes. We close with a polypharmacy case study. A JADER (Tsuchiya et al., 2020) peripheral-neuropathy report documents a patient on FOLFOX (de Gramont et al., 2000) (oxaliplatin, fluorouracil, levofolinate calcium), whose neurotoxicity is dominated by oxaliplatin (Cavaletti & Marmiroli, 2010). Adding a fourth drug requires distinguishing additions that preserve the side effect from those that are not reported with it (Figure 4a). Three candidates co-occur with this core in JADER: *panitumumab*, *bevacizumab*, and *capecitabine*. Panitumumab and bevacizumab look near-identical to a feature-similarity encoder: both are monoclonal antibodies in the same drug-feature cluster (Figure 4b). Yet the data records opposite outcomes: FOLFOX + panitumumab is reported with peripheral neuropathy (label 1, COMP), while FOLFOX + bevacizumab and FOLFOX + capecitabine are not (label 0, INHIB). A feature-similarity predictor cannot resolve this; the contrast is readable only from the joint observation pattern of the 3-drug subset and its 4-drug superset.

For illustration, we train two heads on top of HGPM's pretrained encoder: a relation head over $\{\text{COMP}, \text{EMER}, \text{INHIB}\}$ that types the candidate-regimen interaction, and an adverse-event head over the JADER side-effect vocabulary that ranks the resulting toxicities. The relation head separates the three branches correctly (Figure 4c): COMP$=0.78$ for panitumumab, INHIB$=0.69$ for bevacizumab, INHIB$=0.77$ for capecitabine. Bevacizumab is feature-similar to panitumumab yet flagged as INHIB, read from indicator-pair structure rather than drug embeddings, while the KHGNN baseline collapses both INHIB candidates onto the compositional class (Figure 5, Appendix G). Going beyond the binary class, the adverse-event head (Figure 4d) recovers each candidate's drug-class signature: peripheral neuropathy for panitumumab, hypertension for the anti-VEGF bevacizumab, and hand-foot syndrome for the fluoropyrimidine capecitabine. The baseline collapses all three onto a shared neuropathy prior (Figure 6, Appendix G).

7 Conclusion

We introduced HGPM, a hypergraph learning framework that makes *interaction compositionality* explicit in the representation: the joint pattern of subset presence and absence that distinguishes higher-order relations from pairwise ones. HGPM tokenizes subsets as nodes of an inclusion DAG with composition-labeled edges and trains an inclusion-aware Transformer under a masked reconstruction objective. On eight hypergraph benchmarks and two drug-interaction corpora, HGPM matches or exceeds state-of-the-art methods. Extended discussion is in Appendix A.

AI USE STATEMENT

In this work, we used generative AI tools only for language polishing, improving the grammar, wording, and clarity of text drafted by the authors. We have not used generative AI tools for research ideation, for the design of the proposed method or the experimental protocol, for implementation, for the generation or analysis of results, for the creation of figures and tables, or for literature discovery and citation; all of these are the original work of the authors. We have reviewed all AI-assisted work, checking every suggested revision against the authors' original text and the underlying results, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The construction of the target-centered inclusion DAG, including the observed, sub-hyperedge, and perturbation tokens and the per-order budgets, is given in Section 4 and Appendix C.1; the drug-set instantiation, including which records may provide token labels, is given in Appendix C.3. Dataset processing, including the construction of JADER, is described in Appendix D. The hyperparameter search space and the per-dataset configurations used for the reported results are in Appendix E.1 and Appendix E.4. The theoretical results state their assumptions and give complete proofs in Appendix B. All code and scripts will be released upon publication.

ETHICS STATEMENT

Two of our ten evaluation corpora derive from pharmacovigilance systems that collect reports about real patients: HODDI (Wang et al., 2025c) is an academic benchmark built from the U.S. FDA Adverse Event Reporting System (FAERS), and JADER is the public release of Japan's PMDA adverse-event database, used under PMDA's terms of use. Both are de-identified. We did not collect data from human subjects, performed no re-identification, and linked the records to no other data source. The paper reports no narrative fields, and the qualitative case study identifies records only by drug combination and adverse-event term. Spontaneous reports under-report adverse events (Hazell & Shakir, 2006) and record co-medication rather than exposure. An unrecorded drug set is therefore not evidence that a combination is safe or does not occur. The COMP/EMER/INHIB regimes are defined purely by the observation pattern, inherit these biases, and carry no pharmacological or causal meaning.

HGPM produces statistical prioritizations of drug combinations, not clinical recommendations. An INHIB pattern does not mean that a drug protects against a side effect, and a COMP pattern does not establish a drug interaction. Any downstream use should remain at the stage of hypothesis generation and signal screening, with clinical oversight. A model that ranks drug combinations by predicted adverse-event reporting could in principle be queried for harmful combinations. We judge this risk to be low relative to the public source databases, which already list reported combinations, and we introduce no generative component.

REFERENCES

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning (ICML)*, 2022.

Song Bai, Feihu Zhang, and Philip H. S. Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 2021.

Austin R. Benson, Rediet Abebe, Michael T. Schaub, Ali Jadbabaie, and Jon Kleinberg. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*, 115(48):E11221–E11230, 2018. doi: 10.1073/pnas.1800683115.

Cristian Bodnar, Fabrizio Frasca, Nina Otter, Yu Guang Wang, Pietro Liò, Guido F. Montúfar, and Michael Bronstein. Weisfeiler and lehman go cellular: CW networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Cristian Bodnar, Fabrizio Frasca, Yu Guang Wang, Nina Otter, Guido Montúfar, Pietro Liò, and Michael Bronstein. Weisfeiler and lehman go topological: Message passing simplicial networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021b.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Guido Cavaletti and Paola Marmiroli. Chemotherapy-induced peripheral neurotoxicity. *Nature Reviews Neurology*, 2010.

Yujie Chen, Tengfei Ma, Xixi Yang, Jianmin Wang, Bosheng Song, and Xiangxiang Zeng. MUFFIN: multi-scale feature fusion for drug–drug interaction prediction. *Bioinformatics*, 2021.

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.

Ting-Chao Chou. Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacological Reviews*, 2006.

Aimery de Gramont, A. Figer, M. Seymour, M. Homerin, A. Hmissi, J. Cassidy, C. Boni, H. Cortes-Funes, A. Cervantes, G. Freyer, D. Papamichael, N. Le Bail, C. Louvet, D. Hendler, F. de Braud, C. Wilson, F. Morvan, and A. Bonetti. Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. *Journal of Clinical Oncology*, 2000.

Yihe Dong, Will Sawin, and Yoshua Bengio. HNHN: Hypergraph networks with hyperedge neurons. In *ICML Graph Representation Learning and Beyond (GRL+) Workshop*, 2020.

Iulia Duta, Giulia Cassarà, Fabrizio Silvestri, and Pietro Liò. Sheaf hypergraph networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Stefania Ebli, Michaël Defferrard, and Gard Spreemann. Simplicial neural networks. In *NeurIPS 2020 Workshop on Topological Data Analysis and Beyond*, 2020.

Floor Eijkelboom, Rob Hesselink, and Erik J. Bekkers. E(n) equivariant message passing simplicial networks. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Yue Gao, Yifan Feng, Shiquan Liu, Xiangmin Han, Shaoyi Du, Zongze Wu, and Han Hu. Hypergraph foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4063–4080, 2026. doi: 10.1109/TPAMI.2025.3647504.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017. URL `https://proceedings.mlr.press/v70/gilmer17a.html`.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mustafa Hajij, Ghada Zamzmi, Theodore Papamarkou, Nina Miolane, Aldo Guzmán-Sáenz, Karthikeyan Natesan Ramamurthy, Tolga Birdal, Tamal K. Dey, Soham Mukherjee, Shreyas N. Samaga, Neal Livesay, Robin Walters, Paul Rosen, and Michael T. Schaub. Topological deep learning: Going beyond graph data. *arXiv preprint arXiv:2206.00606*, 2023.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 1963.

Lorna Hazell and Saad A. W. Shakir. Under-reporting of adverse drug reactions: A systematic review. *Drug Safety*, 29(5):385–396, 2006. doi: 10.2165/00002018-200629050-00003.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

Zhenyu Hou, Yufei He, Yukuo Cen, Xiao Liu, Yuxiao Dong, Evgeny Kharlamov, and Jie Tang. GraphMAE2: A decoding-enhanced masked self-supervised graph learner. In *Proceedings of the ACM Web Conference (WWW)*, 2023.

Jing Huang and Jie Yang. UniGNN: a unified framework for graph and hypergraph neural networks. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Hyunjin Hwang, Seungwoo Lee, Chanyoung Park, and Kijung Shin. AHP: Learning to negative sample for hyperedge prediction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 2237–2242, New York, NY, USA, 2022. ACM. doi: 10.1145/3477495.3531836.

Sunwoo Kim, Shinhwan Kang, Fanchen Bu, Soo Yong Lee, Jaemin Yoo, and Kijung Shin. HypeBoy: Generative self-supervised representation learning on hypergraphs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DZUzOKE6og.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser, 2002.

Devin Kreuzer, Dominique Beaini, William L. Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Dongjin Lee and Kijung Shin. I'm me, we're us, and I'm us: Tri-directional contrastive learning on hypergraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8456–8464. Association for the Advancement of Artificial Intelligence (AAAI), 2023. doi: 10.1609/aaai.v37i7.26019.

Ming Li, Yujie Fang, Yi Wang, Han Feng, Yongchun Gu, Lu Bai, and Pietro Liò. Deep hypergraph neural networks with tight framelets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Natalie Ann Lozano-Huntelman, April Zhou, Elif Tekin, Mauricio Cruz-Loya, Bjørn Østman, Sada Boyd, Van M. Savage, and Pamela Yeh. Hidden suppressive interactions are common in higher-order drug combinations. *iScience*, 24(4):102355, 2021. doi: 10.1016/j.isci.2021.102355.

Tianyi Ma, Yiyue Qian, Shinan Zhang, Chuxu Zhang, and Yanfang Ye. Adaptive expansion for hypergraph learning. *arXiv preprint arXiv:2502.15564*, 2025a.

Tianyi Ma, Yiyue Qian, Zheyuan Zhang, Zehong Wang, Shinan Zhang, Chuxu Zhang, and Yanfang Ye. Hypergraph representation learning with adaptive broadcasting and receiving. In *ICDM*, 2025b.

Tianyi Ma, Yiyue Qian, Zehong Wang, Zheyuan Zhang, Chuxu Zhang, and Yanfang Ye. Bhygnn+: Unsupervised representation learning for heterophilic hypergraphs. *arXiv preprint arXiv:2602.14919*, 2026a.

Yijun Ma, Zehong Wang, Weixiang Sun, and Yanfang Ye. Temporal graph pattern machine. *arXiv preprint arXiv:2601.22454*, 2026b.

Boris Mityagin. The zero set of a real analytic function. *Mathematical Notes*, 2020.

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and Leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Duc Anh Nguyen, Canh Hao Nguyen, Peter Petschner, and Hiroshi Mamitsuka. SPARSE: a sparse hypergraph neural network for learning multiple types of latent combinations to accurately predict drug–drug interactions. *Bioinformatics*, 2022.

Arnold K Nyamabo, Hui Yu, and Jian-Yu Shi. SSI-DDI: substructure–substructure interactions for drug–drug interaction prediction. *Briefings in Bioinformatics*, 2021.

Prasanna Patil, Govind Sharma, and M. Narasimha Murty. Negative sampling for hyperlink prediction in networks. In Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan (eds.), *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II*, volume 12085 of *Lecture Notes in Computer Science*, pp. 607–619, Cham, 2020. Springer International Publishing. doi: 10.1007/978-3-030-47436-2_46.

Bo Peng and Xia Ning. Deep learning for high-order drug-drug interaction prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019.

Kristina Preuer, Richard P I Lewis, Sepp Hochreiter, Andreas Bender, Krishna C Bulusu, and Günter Klambauer. DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, 2018.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Gian-Carlo Rota. On the foundations of combinatorial theory I: Theory of Möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1964.

Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences*, 2018.

Khaled Mohammed Saifuddin, Briana Bumgardner, Farhan Tanvir, and Esra Akbas. HyGNN: Drug-drug interaction prediction via hypergraph neural network. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 2023.

Li Sun, Ming Zhang, Wenxin Jin, Zhongtian Sun, Zhenhao Huang, Hao Peng, Sen Su, and Philip S. Yu. Heterophily-agnostic hypergraph neural networks with Riemannian local exchanger. In *Proceedings of the ACM Web Conference (WWW)*, 2026.

Bohan Tang, Siheng Chen, and Xiaowen Dong. Hypergraph-MLP: Learning on hypergraphs without message passing. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Nicholas P. Tatonetti, Patrick P. Ye, Roxana Daneshjou, and Russ B. Altman. Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 2012.

Elif Tekin, Cynthia White, Tina Manzhu Kang, Nina Singh, Mauricio Cruz-Loya, Robert Damoiseaux, Van M. Savage, and Pamela J. Yeh. Prevalence and patterns of higher-order drug interactions in Escherichia coli. *npj Systems Biology and Applications*, 4(1):31, 2018. doi: 10.1038/s41540-018-0069-9.

Lev Telyatnikov, Maria Sofia Bucarelli, Guillermo Bernardez, Olga Zaghen, Simone Scardapane, and Pietro Liò. Hypergraph neural networks through the lens of message passing: A common perspective to homophily and architecture design. *Transactions on Machine Learning Research (TMLR)*, 2025.

Masami Tsuchiya, Taku Obara, Takamasa Sakai, Kaori Nomura, Chizuko Takamura, and Nariyasu Mano. Quality evaluation of the Japanese Adverse Drug Event Report database (JADER). *Pharmacoepidemiology and Drug Safety*, 2020.

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

Peihao Wang, Shenghao Yang, Yunyu Liu, Zhangyang Wang, and Pan Li. Equivariant hypergraph diffusion neural operators. In *International Conference on Learning Representations (ICLR)*, 2023a.

Yuxin Wang, Quan Gan, Xipeng Qiu, Xuanjing Huang, and David Wipf. From hypergraph energy functions to hypergraph neural networks. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023b.

Zehong Wang, Zheyuan Zhang, Nitesh V. Chawla, Chuxu Zhang, and Yanfang Ye. GFT: Graph foundation model with transferable tree vocabulary. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zehong Wang, Zheyuan Zhang, Tianyi Ma, Nitesh V. Chawla, Chuxu Zhang, and Yanfang Ye. Beyond message passing: Neural graph pattern machine. In *International Conference on Machine Learning (ICML)*, 2025a.

Zehong Wang, Zheyuan Zhang, Tianyi Ma, Chuxu Zhang, and Yanfang Ye. Generative graph pattern machine. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Zhaoying Wang, Yingdan Shi, Xiang Liu, Can Chen, Jun Wen, and Ren Wang. HODDI: A dataset of high-order drug-drug interactions for computational pharmacovigilance. *arXiv preprint arXiv:2502.06274*, 2025c.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. Mole-BERT: Rethinking pre-training graph neural networks for molecules. In *International Conference on Learning Representations (ICLR)*, 2023.

Linhuang Xie, Shihao Gao, Jie Liu, Ming Yin, and Taisong Jin. K-hop hypergraph neural network: A comprehensive aggregation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

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

Chaoqi Yang, Ruijie Wang, Shuochao Yao, and Tarek F. Abdelzaher. Semi-supervised hypergraph node classification on hypergraph line expansion. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM)*, 2022.

Jiannan Yang, Zhongzhi Xu, William Ka Kei Wu, Qian Chu, and Qingpeng Zhang. GraphSynergy: a network-inspired deep learning model for anticancer drug combination prediction. *Journal of the American Medical Informatics Association*, 2021.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 2018.

A IMPACT STATEMENT

This work introduces the Hypergraph Pattern Machine, a framework that makes *interaction compositionality* an explicit part of the input for hypergraph learning, distinguishing whether a higher-order relation is compositional, emergent, or inhibitory with respect to its subsets. By exposing the compositionality structure of hypergraph data to the model, our framework aims to provide more interpretable signals for higher-order interaction prediction that can support pharmacovigilance screening, particularly in polypharmacy. As machine learning systems for higher-order relational data are increasingly applied to clinical, biological, and societal contexts, it is important to consider the broader implications of advances in this representational paradigm.

Potential Benefits. HGPM may benefit several scientific and clinical domains. In polypharmacy and pharmacology (Wang et al., 2025c), the explicit COMP/EMER/INHIB regimes provide structurally distinct screening signals that separate reported associations that persist, emerge, or are absent when a drug is added; they describe reporting patterns and require clinical and pharmacological validation before any use. In drug discovery, representations that expose inclusion-hierarchy compositionality may surface higher-order patterns that message-passing models do not represent, complementing existing pairwise drug–drug interaction pipelines. Beyond the drug domain, HGPM applies to any higher-order relational data—recommendation, social-network analysis, and biological interaction modeling—where compositionality between observed and absent subsets carries scientifically meaningful signal. More broadly, our framework contributes a shift in hypergraph learning from message passing on observed structure toward representations that encode the joint pattern of presence and absence, providing a principled interface between hypergraph machine learning and the structural questions practitioners actually care about.

Risks and Misuse Considerations. As with many advances in clinically adjacent machine learning, improved higher-order interaction prediction may lower barriers to misuse. Mispredictions in clinical settings could affect patient care if HGPM outputs are used without physician oversight; training data such as HODDI (Wang et al., 2025c) derive from spontaneous reporting, which under-reports adverse events (Hazell & Shakir, 2006), and HGPM may inherit this bias; in particular, an unrecorded drug set is not evidence that the combination is safe. The same model that flags hazardous polypharmacy combinations could in principle be inverted to identify adversarial combinations with intentionally harmful effects, although our framework does not introduce new generative objectives beyond existing interaction-prediction pipelines. We therefore emphasize that responsible deployment should follow established practices in clinical decision support, regulatory compliance, and dual-use risk assessment, and that any clinical use should retain human oversight at every step of the decision pipeline.

Limitations and Scope. Three limitations remain: (1) the inclusion DAG grows combinatorially with order; HGPM bounds it by per-order sampling, which is approximate, and adaptive or importance-based subset selection is a natural extension. (2) HGPM is per-target: each prediction builds and processes its own DAG, so whole-graph inference on large graphs requires per-node sampling. However, since the learned component is a standard Transformer, existing Transformer inference frameworks apply directly, keeping the practical cost manageable. (3) The composition labels treat every unrecorded subset as absent; in spontaneous-reporting data, absence may reflect under-reporting (Hazell & Shakir, 2006), so HGPM's drug predictions are screening signals rather than causal or clinical conclusions.

Future Directions. Future work may extend HGPM to weighted, signed, or multi-relational hypergraphs, where the binary observation indicator generalizes to richer label spaces. Scaling pretraining to larger higher-order interaction corpora may improve cross-domain generalization, and integrating compositionality-aware representations with pharmacokinetic or mechanistic simulators could bridge data-driven prediction and mechanistic understanding. From a societal perspective, developing evaluation protocols, dataset-governance practices, and clinical-deployment safeguards for higher-order interaction models will be important to ensure that the representational advances introduced here translate into responsible scientific and clinical use.

B THEORETICAL ANALYSIS

This appendix collects the formal results that ground HGPM's design. Section B.1 shows that the composition labels COMP/EMER/INHIB (Section 3) are exactly the values of the discrete forward-difference (cellular coboundary) of the hyperedge indicator on the inclusion lattice, with a corollary characterizing simplicial complexes as the EMER-free stratum. Section B.2 establishes that hyperedge-MP encoders cannot represent the compositionality structure $\mathcal{C}(\mathcal{H})$ when node features are uninformative: there exists a hypergraph pair on which all such encoders collapse to identical features but $\mathcal{C}(\mathcal{H})$ differs. Section B.3 then exhibits a non-isomorphic, WL-equivalent hypergraph pair on which HGPM's centered representation distinguishes the two, generically over weights, under an idealized deterministic tokenizer.

Notational Conventions Used Throughout. $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ is a hypergraph with vertex set $\mathcal{V}$, hyperedge set $\mathcal{E} \subseteq 2^{\mathcal{V}}$, and node features $\mathbf{X}$. The hyperedge indicator is $\mathbb{1}_S = \mathbb{1}[S \in \mathcal{E}]$, with $\mathbb{1}_\emptyset = 0$ by convention ($\emptyset$ is never a hyperedge). The compositionality structure $\mathcal{C}(\mathcal{H})$ collects covering pairs (S, S') with $S \subset S'$, $|S'| = |S| + 1$, and not-both-absent indicators, labeled by comp as defined in Section 3. The set of all covering pairs is denoted

$$\mathrm{Cov}(\mathcal{V}) = \{(S, S') : S \subset S' \subseteq \mathcal{V},\ |S'| = |S| + 1\},$$

and we write $\mathrm{Cov}_+(\mathcal{V}) = \{(S, S') \in \mathrm{Cov}(\mathcal{V}) : S \neq \emptyset\}$ for covers with non-empty bottom face. We use "covering pair" (cover) and "adjacent-order pair" interchangeably throughout this appendix; the latter is the term used in the main text.

B.1 MATHEMATICAL IDENTITY OF COMPOSITIONALITY

The forward-difference (cellular coboundary) operator δ acts on functions $f : 2^{\mathcal{V}} \to \mathbb{R}$ as

$$(\delta f)(S, S') = f(S') - f(S), \qquad (S, S') \in \mathrm{Cov}(\mathcal{V}).$$

Applied to the hyperedge indicator viewed as a function $\mathbb{1}_{(\cdot)} : S \mapsto \mathbb{1}_S$, this yields a $\{-1, 0, +1\}$-valued 1-cochain on the Hasse cover.

Theorem 1 (Coboundary Identity)**.** *For every $(S, S') \in \mathrm{Cov}(\mathcal{V})$:*

(a) (*Local bijection.*) $(\delta\mathbb{1}_{(\cdot)})(S, S') \in \{-1, 0, +1\}$*, and*

$$\begin{aligned} \mathrm{comp}(S, S') = \textsc{Emer} &\iff (\delta\mathbb{1}_{(\cdot)})(S, S') = +1, \\ \mathrm{comp}(S, S') = \textsc{Inhib} &\iff (\delta\mathbb{1}_{(\cdot)})(S, S') = -1, \\ \mathrm{comp}(S, S') = \textsc{Comp} &\iff (\delta\mathbb{1}_{(\cdot)})(S, S') = 0 \text{ and } \mathbb{1}_S = 1, \end{aligned}$$

with both-absent pairs corresponding to $(\delta\mathbb{1}_{(\cdot)})(S, S') = 0$ and $\mathbb{1}_S = 0$.

(b) (*Information equivalence.*) *Given $\mathcal{V}$ and the anchor $\mathbb{1}_\emptyset = 0$, the indicator $\mathbb{1}_{(\cdot)}$ is uniquely recoverable from $\mathcal{C}(\mathcal{H})$ by chain-induction along any saturated chain in $(2^{\mathcal{V}}, \subseteq)$.*

Proof. **(a)** Since $\mathbb{1}_S \in \{0, 1\}$, $(\delta\mathbb{1}_{(\cdot)})(S, S') = \mathbb{1}_{S'} - \mathbb{1}_S \in \{-1, 0, +1\}$. The four ordered pairs $(\mathbb{1}_S, \mathbb{1}_{S'}) \in \{0, 1\}^2$ map bijectively to $((\delta\mathbb{1}_{(\cdot)})(S, S'),\ \mathbb{1}_S)$ as: $(0, 1) \mapsto (+1, 0)$, $(1, 0) \mapsto (-1, 1)$, $(1, 1) \mapsto (0, 1)$, $(0, 0) \mapsto (0, 0)$. The biconditionals follow from the definition of comp in Section 3.

(b) For any $S \subseteq \mathcal{V}$, choose a saturated chain $\emptyset = S_0 \subset S_1 \subset \cdots \subset S_n = S$. Telescoping with $\mathbb{1}_\emptyset = 0$,

$$\mathbb{1}_S = \sum_{i=1}^{n} (\delta\mathbb{1}_{(\cdot)})(S_{i-1}, S_i).$$

By Clause (a), each summand is determined by $\mathcal{C}(\mathcal{H})$: if the cover (S_{i-1}, S_i) is informative, its label gives the value; if both indicators are zero, the value is zero (and $\mathbb{1}_{S_{i-1}} = 0$ holds inductively, since otherwise the cover would be either COMP or INHIB and hence labeled in $\mathcal{C}(\mathcal{H})$). The right-hand sum is path-independent because $\delta\mathbb{1}_{(\cdot)}$ is exact: it is the coboundary of a 0-cochain. □

Corollary 1 (Simplicial complexes are exactly the no-EMER stratum above the empty face)**.** *$\mathcal{E}$ is downward-closed among non-empty subsets — $\sigma \in \mathcal{E}$ and $\emptyset \neq \tau \subset \sigma$ imply $\tau \in \mathcal{E}$, equivalently $\mathcal{H}$ is an abstract simplicial complex — if and only if $(\delta\mathbb{1}_{(\cdot)})(S, S') \neq +1$ for every $(S, S') \in \mathrm{Cov}_+(\mathcal{V})$. Equivalently, $\mathcal{C}(\mathcal{H})$ contains no* EMER *label on covers with $|S| \geq 1$.*

Proof. ($\Rightarrow$) Let $\mathcal{E}$ be downward-closed on non-empty subsets and fix $(S, S') \in \mathrm{Cov}_+(\mathcal{V})$. If $(\delta\mathbb{1}_{(\cdot)})(S, S') = +1$, then $\mathbb{1}_{S'} = 1, \mathbb{1}_S = 0$. Since $S \neq \emptyset$ and $S \subset S'$, downward closure forces $S \in \mathcal{E}$, contradicting $\mathbb{1}_S = 0$.

($\Leftarrow$) Suppose $(\delta\mathbb{1}_{(\cdot)}) \neq +1$ on $\mathrm{Cov}_+(\mathcal{V})$. Fix $\sigma \in \mathcal{E}$ and non-empty $\tau \subset \sigma$; set $k = |\sigma| - |\tau| \geq 1$. Choose a saturated chain $\tau = T_0 \subset T_1 \subset \cdots \subset T_k = \sigma$. Each $T_{i-1} \supseteq \tau \neq \emptyset$, so $(T_{i-1}, T_i) \in \mathrm{Cov}_+(\mathcal{V})$. Telescoping,

$$\mathbb{1}_\sigma - \mathbb{1}_\tau = \sum_{i=1}^{k} (\delta\mathbb{1}_{(\cdot)})(T_{i-1}, T_i) \leq 0,$$

since each summand is in $\{-1, 0\}$ by hypothesis. Hence $\mathbb{1}_\tau \geq \mathbb{1}_\sigma = 1$, so $\mathbb{1}_\tau = 1$ (as $\mathbb{1}_{(\cdot)} \in \{0, 1\}$), i.e., $\tau \in \mathcal{E}$. □

Remark 1 (Why the empty bottom face is excluded). The restriction to $\mathrm{Cov}_+(\mathcal{V})$ is forced by the convention $\mathbb{1}_\emptyset = 0$. For any $\mathcal{E} \neq \emptyset$ and any $\{v\} \in \mathcal{E}$, the cover $(\emptyset, \{v\})$ is automatically EMER ($\mathbb{1}_\emptyset = 0, \mathbb{1}_{\{v\}} = 1$); this bottom-layer EMER carries no information about closure structure. Concretely, $\mathcal{V} = \{1, 2, 3\}$ with $\mathcal{E} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$ is downward-closed among non-empty subsets, yet $(\emptyset, \{i\})$ is EMER for each i. On $\mathrm{Cov}_+(\mathcal{V})$, all nine remaining covers have $(\delta\mathbb{1}_{(\cdot)}) = 0$, consistent with Corollary 1.

Remark 2 (Implication for the simplicial-complex contrast). Corollary 1 shows that simplicial-complex networks foreclose emergent transitions on non-empty faces by construction; inhibitory transitions (a face without a coface, e.g., $(\{1, 2\}, \{1, 2, 3\})$ in the complex $\{1\}, \{2\}, \{3\}, \{1, 2\}$) remain possible. Any abstract simplicial complex satisfies

$$\big\{\mathbb{1}_\emptyset = 0,\ (\delta\mathbb{1}_{(\cdot)})(S, S') \neq +1\ \forall (S, S') \in \mathrm{Cov}_+(\mathcal{V})\big\},$$

the EMER-free stratum on $\mathrm{Cov}_+(\mathcal{V})$. Simplicial-complex encoders therefore cannot, by virtue of their input domain, represent or distinguish EMER regimes on covers with non-empty bottom face — i.e., on any genuine "two existing faces of consecutive size" transition. HGPM's input admits arbitrary indicator patterns on the sampled covers, which its attention reads through the role-pair and direction biases jointly.

Remark 3 (Connection to Möbius inversion and Harsanyi dividends). The Möbius transform of $\mathbb{1}_{(\cdot)}$ on $(2^\mathcal{V}, \subseteq)$ is the unique $g : 2^\mathcal{V} \to \mathbb{Z}$ satisfying $\mathbb{1}_S = \sum_{T \subseteq S} g(T)$, equivalently $g(S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} \mathbb{1}_T$ (Rota, 1964; Harsanyi, 1963). By Theorem 1(b), $\mathbb{1}_{(\cdot)}$ — and therefore g — is recoverable from $\mathcal{C}(\mathcal{H})$ given $\mathcal{V}$ and the anchor; this places $\mathcal{C}(\mathcal{H})$ inside the classical Möbius / Harsanyi framework on Boolean lattices. We use the connection conceptually only; Möbius coefficients are not computed by HGPM.

A quantitative caveat: "$g(S) \neq 0 \Leftrightarrow$ EMER/INHIB pair on a chain in 2^S" fails by cancellation. Take $\mathcal{V} = \{a, b, c\}$, $\mathcal{E} = \{\{a\}, \{a, b\}\}$. Direct computation of all eight terms yields

$$g(\mathcal{V}) = 0 - 1 - 0 - 0 + 1 + 0 + 0 - 0 = 0,$$

yet $2^\mathcal{V}$ contains EMER covers $(\emptyset, \{a\})$ and $(\{b\}, \{a, b\})$, and INHIB covers $(\{a\}, \{a, c\})$ and $(\{a, b\}, \{a, b, c\})$. Thus $g(S) \neq 0$ implies non-constant $\mathbb{1}_{(\cdot)}$ on 2^S (one direction); the converse can fail.

B.2 Hyperedge Message Passing Cannot Represent Compositionality

Definition 1 (Hyperedge-MP encoder)**.** A *hyperedge-MP encoder* of depth T is any architecture of the form

$$\begin{aligned}
\mathbf{m}_e^{(t)} &= \phi_E^{(t)}\big(\mathbf{h}_e^{(t)},\ \{\!\{\mathbf{h}_v^{(t)} : v \in e\}\!\}\big),\\
\mathbf{h}_v^{(t+1)} &= \phi_V^{(t)}\big(\mathbf{h}_v^{(t)},\ \{\!\{\mathbf{m}_e^{(t)} : v \in e\}\!\}\big),\\
\mathbf{h}_e^{(t+1)} &= \psi^{(t)}\big(\mathbf{m}_e^{(t)},\ \{\!\{\mathbf{h}_v^{(t+1)} : v \in e\}\!\}\big),
\end{aligned}$$

with $\mathbf{h}_v^{(0)} = \mathbf{x}_v$ and $\mathbf{h}_e^{(0)}$ a fixed initialization token, where each $\phi_E^{(t)}, \phi_V^{(t)}, \psi^{(t)}$ is a deterministic function of (own feature, multiset of incident features). HGNN (Feng et al., 2019), AllDeepSets / AllSetTransformer (Chien et al., 2022), ED-HNN (Wang et al., 2023a), and UniGCN (Huang & Yang, 2021) all admit this factorization.

Lemma 1 (Bipartite-WL upper bound on hyperedge-MP)**.** *Let $B(\mathcal{H}) = (\mathcal{V} \sqcup \mathcal{E}, I)$, $I = \{(v, e) : v \in e\}$, denote the incidence bipartite graph of $\mathcal{H}$. Bipartite-WL refines colors by*

$$c_v^{(t+1)} = \mathrm{HASH}\big(c_v^{(t)},\ \{\!\{c_e^{(t)} : v \in e\}\!\}\big), \quad c_e^{(t+1)} = \mathrm{HASH}\big(c_e^{(t)},\ \{\!\{c_v^{(t)} : v \in e\}\!\}\big).$$

For any hyperedge-MP encoder Φ of depth T and any two hypergraphs $\mathcal{H}_1, \mathcal{H}_2$ on the same vertex set $\mathcal{V}$, if bipartite-WL produces, at every iteration t, pointwise-equal vertex colors $c_v^{(t)}[\mathcal{H}_1] = c_v^{(t)}[\mathcal{H}_2]$ for all $v \in \mathcal{V}$ together with pointwise-equal per-vertex multisets of incident hyperedge colors, then

$$\mathbf{h}_v^{(T)}[\mathcal{H}_1] = \mathbf{h}_v^{(T)}[\mathcal{H}_2]\ \forall v \in \mathcal{V}, \qquad \{\!\{\mathbf{h}_e^{(T)}[\mathcal{H}_1]\}\!\} = \{\!\{\mathbf{h}_e^{(T)}[\mathcal{H}_2]\}\!\}.$$

Proof. Induction on t. At $t = 0$, $\mathbf{h}_v^{(0)} = \mathbf{x}_v$ depends only on $c_v^{(0)} = \mathbf{x}_v$; $\mathbf{h}_e^{(0)}$ is a fixed token, identical across $\mathcal{H}_1, \mathcal{H}_2$. Inductive step: assume at iteration t both vertex features and per-vertex incident-hyperedge feature multisets agree across $\mathcal{H}_1, \mathcal{H}_2$. Then both inputs to $\phi_V^{(t)}$ at every v are identical, so $\mathbf{h}_v^{(t+1)}[\mathcal{H}_1] = \mathbf{h}_v^{(t+1)}[\mathcal{H}_2]$. The hyperedge update is symmetric. This is the standard GIN-WL bound (Xu et al., 2019) lifted to bipartite incidence; see also Huang & Yang (2021, Prop. 2), which bounds UniGNN by the one-dimensional generalized WL test. □

Theorem 2 (Message-Passing Blindness)**.** *There exist hypergraphs $\mathcal{H}_1 = (\mathcal{V}, \mathcal{E}_1, \mathbf{X})$ and $\mathcal{H}_2 = (\mathcal{V}, \mathcal{E}_2, \mathbf{X})$ over the same vertex set $\mathcal{V}$ and constant node features $\mathbf{x}_v = \mathbf{x}_0$ such that:*

(a) For every hyperedge-MP encoder Φ of any finite depth T,

$$\mathbf{h}_v^{(T)}[\mathcal{H}_1] = \mathbf{h}_v^{(T)}[\mathcal{H}_2]\ \forall v \in \mathcal{V}, \quad \{\!\{\mathbf{h}_e^{(T)}[\mathcal{H}_1]\}\!\} = \{\!\{\mathbf{h}_e^{(T)}[\mathcal{H}_2]\}\!\}.$$

(b) $\mathcal{C}(\mathcal{H}_1) \neq \mathcal{C}(\mathcal{H}_2)$ as labeled sets of adjacent-order pairs.

Consequently, every map f whose only $\mathcal{H}$-dependent input is the encoder output multisets satisfies $f(\mathcal{H}_1) = f(\mathcal{H}_2)$, hence cannot recover $\mathcal{C}(\mathcal{H})$ on at least one of $\mathcal{H}_1, \mathcal{H}_2$.

Proof. **Construction.** $\mathcal{V} = \{1, 2, 3, 4, 5, 6\}$, $\mathbf{x}_v = \mathbf{x}_0$ constant, $\mathcal{E}_1 = \{\{1, 2, 3\}, \{4, 5, 6\}\}$, $\mathcal{E}_2 = \{\{1, 2, 4\}, \{3, 5, 6\}\}$.

Part (a) — bipartite-WL collapse. Both $B(\mathcal{H}_1)$ and $B(\mathcal{H}_2)$ are disjoint unions of two stars $K_{1,3}$. Hand-running WL with constant initialization:

- $t = 0$: $c_v^{(0)} = A$ for all v; $c_e^{(0)} = \perp$ for both hyperedges.
- $t = 1$: every vertex is in exactly one hyperedge, so $c_v^{(1)} = \mathrm{HASH}(A, \{\!\{\perp\}\!\}) = A_1$ uniformly; every hyperedge sees $\{\!\{A, A, A\}\!\}$, so $c_e^{(1)} = \mathrm{HASH}(\perp, \{\!\{A, A, A\}\!\}) = B_1$ uniformly.
- $t = 2$: $c_v^{(2)} = A_2$, $c_e^{(2)} = B_2$ uniformly. Stable.

Color multisets are pointwise identical on $\mathcal{H}_1, \mathcal{H}_2$, and the per-vertex multiset of incident hyperedge colors equals $\{\!\{B_t\}\!\}$ for both. Lemma 1 yields equal encoder outputs.

Part (b) — compositionality differs. Take $(S, S') = (\{1, 2\}, \{1, 2, 3\})$. In $\mathcal{H}_1$: $\mathbb{1}_{\{1,2\}} = 0$ and $\mathbb{1}_{\{1,2,3\}} = 1$, so $\mathrm{comp}(S, S') = \textsc{Emer}$ and $((S, S'), \textsc{Emer}) \in \mathcal{C}(\mathcal{H}_1)$. In $\mathcal{H}_2$: $\{1, 2, 3\} \notin \mathcal{E}_2$, so the cover is both-absent and excluded from $\mathcal{C}(\mathcal{H}_2)$. Hence $\mathcal{C}(\mathcal{H}_1) \neq \mathcal{C}(\mathcal{H}_2)$.

Consequence. Any f depending on $\mathcal{H}$ only through encoder outputs gives $f(\mathcal{H}_1) = f(\mathcal{H}_2)$ by (a). Since $\mathcal{C}$ differs at the named pair (S, S'), f errs on at least one of $\mathcal{H}_1, \mathcal{H}_2$. □

Remark 4 (Representation expressiveness vs. end-to-end predictability). Theorem 2 is a representation-level claim: any predictor confined to encoder outputs errs on at least one of $\mathcal{H}_1, \mathcal{H}_2$. It does not preclude an end-to-end pipeline in which a head reads the raw incidence list $\mathcal{E}$ and computes $\mathbb{1}_{S \in \mathcal{E}}$ by membership lookup, bypassing the encoder. Such a lookup-head trivially solves compositionality but renders the encoder dead weight. HGPM (Section 4.1) lifts this barrier by exposing $\mathbb{1}_S$ as a token attribute and the inclusion relation as an attention bias, so the encoder itself becomes the carrier of the compositionality signal.

Remark 5 (On the constant-features assumption). The construction relies on $\mathbf{x}_v = \mathbf{x}_0$ for the WL collapse; with sufficiently distinguishing features, bipartite-WL separates $\mathcal{H}_1$ from $\mathcal{H}_2$ at iteration 1. The hypergraphs $\mathcal{H}_1, \mathcal{H}_2$ are also isomorphic as unlabeled hypergraphs (witness: the transposition $3 \leftrightarrow 4$). This is the standard regime for WL-style separation results in graph learning (Xu et al., 2019; Morris et al., 2019): a featureless minimal counterexample establishes non-universality. Because the pair is isomorphic and vertex-transitive, HGPM (with the idealized tokenizer of Lemma 3, or with $\mathbf{e}^{\text{pos}} = 0$) cannot separate it either; it serves as motivation only. Theorem 3 below provides a non-isomorphic witness pair on which Theorem 2 continues to hold.

B.3 HGPM Strictly Separates a WL-Equivalent Pair

We now exhibit a *non-isomorphic* hypergraph pair on which all hyperedge-MP encoders collapse but HGPM's centered representation distinguishes the two, generically over weights. The non-isomorphism avoids the unlabeled-isomorphism degeneracy of Theorem 2. Throughout this section we use the idealized tokenizer of Lemma 3 and restrict to the parameter subspace $\Theta_0 = \{\theta : \mathbf{e}^{\text{pos}} = 0\}$, on which HGPM does not depend on the node-ID-based token order.

Construction 1 (Non-isomorphic WL-equivalent witness pair)**.** Let $\mathcal{V}' = \{1, \dots, 6\}$ with constant features $\mathbf{x}_v = \mathbf{x}_0$, and define

$$\mathcal{E}'_1 = \big\{\{1,2,3\},\ \{1,2,4\},\ \{3,5,6\},\ \{4,5,6\}\big\},$$

$$\mathcal{E}'_2 = \big\{\{1,2,3\},\ \{1,4,5\},\ \{2,4,6\},\ \{3,5,6\}\big\}.$$

Both hypergraphs are 3-uniform with 4 hyperedges; every $v \in \mathcal{V}'$ has degree 2. We refer to the resulting hypergraphs as $\mathcal{H}'_1, \mathcal{H}'_2$.

Verified Structural Properties (Exhaustive Enumeration).

- Pair-intersection multisets: $\{|e \cap e'| : e, e' \in \mathcal{E}'_1, e \neq e'\} = \{0,0,1,1,2,2\}$ versus $\{|e \cap e'| : e, e' \in \mathcal{E}'_2, e \neq e'\} = \{1,1,1,1,1,1\}$. Pair-intersection histogram is an isomorphism invariant, hence $\mathcal{H}'_1 \ncong \mathcal{H}'_2$. (Brute-force check over all 720 vertex permutations confirms no isomorphism exists.)
- Bipartite-WL stable colorings on $\mathcal{H}'_1$ and $\mathcal{H}'_2$ are pointwise identical (Lemma 2 below).

Lemma 2 (Bipartite-WL collapse on Construction 1)**.** *On both $\mathcal{H}'_1$ and $\mathcal{H}'_2$, bipartite-WL stabilizes at iteration 2 with a single $\mathcal{V}'$-side color and a single $\mathcal{E}'$-side color. Consequently, by Lemma 1, every hyperedge-MP encoder Φ produces identical per-vertex features and identical hyperedge-feature multisets on $\mathcal{H}'_1, \mathcal{H}'_2$.*

Proof. At $t = 0$ all vertices have color A (constant features) and all hyperedges color $B = \text{HASH}(\text{size} = 3, \{\!\{A, A, A\}\!\})$. At $t = 1$, every vertex is in exactly two hyperedges of color B and updates to $A_1 = \text{HASH}(A, \{\!\{B, B\}\!\})$; every hyperedge updates to $B_1 = \text{HASH}(B, \{\!\{A, A, A\}\!\})$. Stable: $t \geq 2$ leaves all colors unchanged. Identical on $\mathcal{H}'_1$ and $\mathcal{H}'_2$. □

Let $V^* = \{1, 2, 5, 6\} \subset \mathcal{V}'$. The remaining two targets $\{3, 4\}$ have isomorphic inclusion DAGs across $\mathcal{H}'_1, \mathcal{H}'_2$: at $c = 3$ the two observed hyperedges containing c coincide, and at $c = 4$ both pairs of observed hyperedges meet only in c.

Lemma 3 (Token cardinality differs at V^*)**.** *For target $c \in V^*$, let $\mathcal{S}_c[\mathcal{H}]$ denote the centered token set under an idealized deterministic tokenizer (not the stochastic sampler of Appendix C.1) that retains the center singleton $\{c\}$, all observed hyperedges containing c, and all 2-subsets of those observed hyperedges that contain c (the latter are classified as $\tau = \text{neg}$ tokens with $\mathbb{1}_S = 0$). Then for every $c \in V^*$,*

$$|\mathcal{S}_c[\mathcal{H}'_1]| = 6, \qquad |\mathcal{S}_c[\mathcal{H}'_2]| = 7,$$

and the corresponding token attribute multisets $\{\!\{(|S|, \mathbb{1}_S, \tau(S))\}\!\}$ differ.

Proof. The structural driver: the two observed hyperedges in $\mathcal{H}'_1$ that contain $c \in V^*$ form a 2-overlap pair (sharing two vertices including c), whereas in $\mathcal{H}'_2$ they form a 1-overlap pair (sharing only c).

- $c = 1$: $\mathcal{H}'_1$ obs $\{1,2,3\},\{1,2,4\}$ (overlap $\{1,2\}$), 2-subsets containing 1: $\{1,2\},\{1,3\},\{1,4\}$ (three distinct). $\mathcal{H}'_2$ obs $\{1,2,3\},\{1,4,5\}$ (overlap $\{1\}$), 2-subsets: $\{1,2\},\{1,3\},\{1,4\},\{1,5\}$ (four distinct).
- $c = 2$: $\mathcal{H}'_1$ obs $\{1,2,3\},\{1,2,4\}$, 2-subs $\{1,2\},\{2,3\},\{2,4\}$. $\mathcal{H}'_2$ obs $\{1,2,3\},\{2,4,6\}$, 2-subs $\{1,2\},\{2,3\},\{2,4\},\{2,6\}$.
- $c = 5$: $\mathcal{H}'_1$ obs $\{3,5,6\},\{4,5,6\}$, 2-subs $\{3,5\},\{5,6\},\{4,5\}$. $\mathcal{H}'_2$ obs $\{1,4,5\},\{3,5,6\}$, 2-subs $\{1,5\},\{4,5\},\{3,5\},\{5,6\}$.
- $c = 6$: $\mathcal{H}'_1$ obs $\{3,5,6\},\{4,5,6\}$, 2-subs $\{3,6\},\{5,6\},\{4,6\}$. $\mathcal{H}'_2$ obs $\{2,4,6\},\{3,5,6\}$, 2-subs $\{2,6\},\{4,6\},\{3,6\},\{5,6\}$.

In all four cases $|\mathcal{S}_c[\mathcal{H}'_1]| = 1+2+3 = 6$ and $|\mathcal{S}_c[\mathcal{H}'_2]| = 1+2+4 = 7$, with attribute multisets differing in the count of $(|S|, \mathbb{1}_S, \tau(S)) = (2, 0, \mathrm{neg})$ tokens (3 versus 4). □

Lemma 4 (Inclusion DAGs are non-isomorphic at $c \in V^*$)**.** *For every $c \in V^*$, the inclusion DAGs $\mathcal{G}_c[\mathcal{H}'_1] = (\mathcal{S}_c[\mathcal{H}'_1], \mathcal{R}_c[\mathcal{H}'_1])$ and $\mathcal{G}_c[\mathcal{H}'_2]$ are non-isomorphic as labeled DAGs.*

Proof. We show the label-restricted out-degree multiset $\{\!\{\deg^+ S : S \in \mathcal{S}_c, \tau(S) = \mathrm{neg}\}\!\}$ differs across $\mathcal{H}'_1, \mathcal{H}'_2$. The out-degree of a neg token equals the number of observed hyperedges containing c that have it as a 2-subset.

In $\mathcal{H}'_1$ at any $c \in V^*$: by the 2-overlap pair structure, exactly one neg token is a 2-subset of *both* observed hyperedges (the shared 2-overlap; e.g., $\{1,2\}$ at $c = 1$); the other two neg tokens are subsets of one observed hyperedge each. Out-degree multiset: $\{2,1,1\}$.

In $\mathcal{H}'_2$ at any $c \in V^*$: by the 1-overlap pair structure, every neg token is a 2-subset of exactly one observed hyperedge. Out-degree multiset: $\{1,1,1,1\}$.

The label-restricted out-degree multiset is invariant under any DAG isomorphism preserving τ. Multisets $\{2,1,1\} \neq \{1,1,1,1\}$ (differ in cardinality 3 vs 4 and maximum value 2 vs 1), hence $\mathcal{G}_c[\mathcal{H}'_1] \ncong \mathcal{G}_c[\mathcal{H}'_2]$ for every $c \in V^*$. □

Lemma 5 (Generic separation on non-isomorphic labeled DAGs)**.** *Let $\Theta = \prod_i \mathbb{R}^{d_i}$ be HGPM's parameter space (attribute embeddings $\mathbf{e}^{\mathrm{ord}}, \mathbf{e}^{\mathrm{ex}}, \mathbf{e}^{\mathrm{src}}, \mathbf{e}^{\mathrm{view}}, \mathbf{e}^{\mathrm{pos}}$, the feature projector W_{x}, structural-bias parameters $b^{\mathrm{dir}}, b^{\mathrm{ord}}, b^{\mathrm{ovlp}}, b^{\mathrm{sib}}, b^{\mathrm{comp}}$, per-layer attention $\{W_Q^{(\ell)}, W_K^{(\ell)}, W_V^{(\ell)}, W_O^{(\ell)}\}$, FFN/LayerNorm weights, attention-pool weights $(\mathbf{w}, \mathbf{W})$). For two finite labeled DAGs $\mathcal{G}_1, \mathcal{G}_2$, define $f_i(\theta) = \mathrm{HGPM}_c(\theta)(\mathcal{G}_i)$. Assume:*

(A1) For every fixed value of the first-layer weights of W_{x} (whose ReLU is the only non-analytic operation), f_1, f_2 are real-analytic in the remaining parameters (closed under finite compositions of softmax, $\tanh$, GELU, LayerNorm with $\epsilon > 0$, linear maps).

(A3) For every such fixed value, there exists θ_0 with $f_1(\theta_0) \neq f_2(\theta_0)$.

Then the equality locus $\Theta_= = \{\theta : f_1(\theta) = f_2(\theta)\}$ is Lebesgue-null.

Proof. $g(\theta) = f_1(\theta) - f_2(\theta) \in \mathbb{R}^d$ is real-analytic on each slice by (A1). $\Theta_= = g^{-1}(0) = \bigcap_{j=1}^d g_j^{-1}(0)$. By (A3), at least one component g_j is not identically zero. The zero set of a non-zero real-analytic function on a connected open subset of $\mathbb{R}^N$ has Lebesgue measure zero (Mityagin, 2020; Krantz & Parks, 2002). Hence $\Theta_= \subseteq g_j^{-1}(0)$ has measure zero within each slice of fixed first-layer projector weights, and therefore in Θ by Fubini's theorem. □

For the statement below we write $\mathbf{h}_c^{\mathrm{ctx}}[\theta, \mathcal{H}] := \sum_{S \in \mathcal{S}_c[\mathcal{H}]} \alpha_S \mathbf{h}_S^{(L)}$ for the attention-pooled context branch of the readout in Appendix C.3, viewed as a function of HGPM weights θ; this is the second concatenand of $\mathbf{h}_c^{\mathrm{node}}$, and a separation in $\mathbf{h}_c^{\mathrm{ctx}}$ lifts to a separation in $\mathbf{h}_c^{\mathrm{node}}$ because a concatenation differs whenever one of its blocks differs.

Theorem 3 (HGPM Separation)**.** *Consider Construction 1.*

(a) Every hyperedge-MP encoder produces identical per-vertex features and identical hyperedge-feature multisets on $\mathcal{H}'_1, \mathcal{H}'_2$ (Lemma 2).

(b) For every $c \in V^* = \{1, 2, 5, 6\}$, the inclusion DAGs satisfy $\mathcal{G}_c[\mathcal{H}'_1] \not\cong \mathcal{G}_c[\mathcal{H}'_2]$ (Lemma 4). Consequently, for any HGPM parameterization $\theta \in \Theta_0$ outside a Lebesgue-null set $\Theta^c_=$, the centered representation $\mathbf{h}^{\mathrm{ctx}}_c[\theta, \mathcal{H}'_1]$ differs from $\mathbf{h}^{\mathrm{ctx}}_c[\theta, \mathcal{H}'_2]$.

(c) Compositionality structures differ: $\mathcal{C}(\mathcal{H}'_1) \neq \mathcal{C}(\mathcal{H}'_2)$. The labeled covering pair $((\{1, 2\}, \{1, 2, 4\}), \textsc{Emer})$ lies in $\mathcal{C}(\mathcal{H}'_1) \setminus \mathcal{C}(\mathcal{H}'_2)$.

Proof. **(a)** By Lemma 2.

(b) Fix $c \in V^*$. Lemma 4 gives $\mathcal{G}_c[\mathcal{H}'_1] \not\cong \mathcal{G}_c[\mathcal{H}'_2]$. We construct an explicit separating witness $\theta_0 \in \Theta$ to verify (A3) of Lemma 5.

Witness construction. Choose:

- Embeddings $\mathbf{e}^{\mathrm{ord}} = \mathbf{e}^{\mathrm{ex}} = \mathbf{e}^{\mathrm{view}} = \mathbf{e}^{\mathrm{pos}} = 0$, and $\mathbf{e}^{\mathrm{src}}$ equal to the standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ for $\{\mathrm{center}, \mathrm{obs}, \mathrm{neg}\}$ in $\mathbb{R}^d$ with $d \geq 4$; the second layer of W_{x} is zero, so $W_{\mathrm{x}}(\mathbf{x}_0) = 0$ for any first-layer weights; LayerNorm gains are one and biases zero.
- Structural biases zero: $b^{\mathrm{dir}} = b^{\mathrm{ord}} = b^{\mathrm{ovlp}} = b^{\mathrm{sib}} = b^{\mathrm{comp}} = 0$.
- $W_Q^{(\ell)} = W_K^{(\ell)} = 0$ (uniform attention weights $1/|\mathcal{S}_c|$); $W_V^{(\ell)} = 0$ (zero attention contribution); $W_O^{(\ell)}$ arbitrary; FFN output weight and bias zero.
- Attention-pool weights $\mathbf{w} = \mathbf{W} = 0$ (uniform mean pool $\alpha_S = 1/|\mathcal{S}_c|$).

Under this θ_0 the residual stream is unchanged at every layer ($W_V = 0$ kills attention output and the FFN outputs zero), so $\mathbf{h}_S^{(L)}$ is determined by $\tau(S)$ alone modulo a deterministic per-token LayerNorm. Define v_τ as the common feature of all type-τ tokens (well-defined because all type-τ inputs are equal under θ_0). The pooled output is

$$\mathbf{h}^{\mathrm{ctx}}_c[\mathcal{H}] = \frac{1}{|\mathcal{S}_c[\mathcal{H}]|} \sum_{\tau \in \{\mathrm{center}, \mathrm{neg}, \mathrm{obs}\}} n_\tau[\mathcal{H}]\, v_\tau.$$

By Lemma 3, for $c \in V^*$ the type counts are $(n_{\mathrm{center}}, n_{\mathrm{neg}}, n_{\mathrm{obs}}) = (1, 3, 2)$ in $\mathcal{H}'_1$ (cardinality 6) and $(1, 4, 2)$ in $\mathcal{H}'_2$ (cardinality 7). Subtracting,

$$\mathbf{h}^{\mathrm{ctx}}_c[\mathcal{H}'_1] - \mathbf{h}^{\mathrm{ctx}}_c[\mathcal{H}'_2] = \tfrac{1}{42}\big(v_{\mathrm{center}} - 3\, v_{\mathrm{neg}} + 2\, v_{\mathrm{obs}}\big),$$

using $\frac{1}{6} - \frac{1}{7} = \frac{1}{42}$, $\frac{3}{6} - \frac{4}{7} = -\frac{3}{42}$, $\frac{2}{6} - \frac{2}{7} = \frac{2}{42}$. For $d \geq 4$, $\mathrm{LN}(\mathbf{e}_k) = (\mathbf{e}_k - \frac{1}{d}\mathbf{1})/\sigma$; in any vanishing combination $\sum_k a_k \mathrm{LN}(\mathbf{e}_k) = 0$, coordinate 4 forces $\sum_k a_k = 0$, after which each $a_k = 0$, so $\{v_{\mathrm{center}}, v_{\mathrm{neg}}, v_{\mathrm{obs}}\}$ are linearly independent. The coefficients $(1, -3, 2)$ are not all zero, hence $\mathbf{h}^{\mathrm{ctx}}_c[\mathcal{H}'_1] \neq \mathbf{h}^{\mathrm{ctx}}_c[\mathcal{H}'_2]$, establishing (A3).

(A1) holds because, once the first-layer projector weights are fixed, softmax / $\tanh$ / GELU / LayerNorm / linear maps are real-analytic and finite compositions thereof are real-analytic. The witness is valid for every value of those weights, so (A3) holds on every slice. By Lemma 5 applied on Θ_0, the equality locus $\Theta^c_=$ is Lebesgue-null.

(c) By Theorem 1(b), $\mathcal{E}'_1 \neq \mathcal{E}'_2$ implies $\mathcal{C}(\mathcal{H}'_1) \neq \mathcal{C}(\mathcal{H}'_2)$. Direct verification: $\{1, 2\} \notin \mathcal{E}'_1$ and $\{1, 2, 4\} \in \mathcal{E}'_1$ give EMER in $\mathcal{H}'_1$; $\{1, 2, 4\} \notin \mathcal{E}'_2$ excludes the cover from $\mathcal{C}(\mathcal{H}'_2)$. □

Remark 6 (One-way separation, not strict containment). Theorem 3 establishes a one-way separation on the compositionality task: there is a hypergraph pair on which HGPM's representation distinguishes what hyperedge-MP cannot. It does *not* establish strict containment of expressivity classes. HGPM is per-target with budget-bounded subset sampling and cannot natively simulate hyperedge-MP's cross-vertex global propagation. The two function classes are *incomparable*; HGPM's design buys expressivity on inclusion-hierarchy patterns at the cost of unrestricted graph propagation. Moreover, the witness in Theorem 3 separates through the numbers of tokens of each type; the result shows that HGPM's input is not bounded by bipartite WL, not that the structural biases are necessary for this pair.

Remark 7 (Targets $c \in \{3, 4\}$ are sampling-deficient on Construction 1). At $c \in \{3, 4\}$ the centered token sets $\mathcal{S}_c[\mathcal{H}'_1]$ and $\mathcal{S}_c[\mathcal{H}'_2]$ yield isomorphic inclusion DAGs (at $c = 3$ the observed hyperedges containing c coincide; at $c = 4$ both pairs meet only in c). HGPM's deterministic sampling at these targets is therefore blind to the difference; this is a property of the construction rather than a defect of HGPM. For tasks aggregating across $c \in \mathcal{V}'$ (e.g., graph classification), the four targets in V^* already yield an output difference, so the global task is solvable on this construction by HGPM but not by hyperedge-MP.

Remark 8 (Generic vs. trained weights). "Generic" in Theorem 3(b) means Lebesgue-a.e. in Θ_0. We make no formal claim about trained weights. The ReLU in the feature projector is handled by the slice-wise argument in Lemma 5.

C Technical Details

We collect the implementation specifics deferred from Section 4, organized along the pipeline: building the inclusion DAG (Section C.1), the encoder and pretraining heads on top of it (Section C.2), and the finetuning side, both transfer and downstream readouts (Section C.3).

C.1 Inclusion DAG Construction

Observed and sub-hyperedge tokens. For each center c and each view, every visible hyperedge $e \ni c$ with $2 \le |e| \le K_{\max}$ is a candidate obs token. In addition, for each incident hyperedge e (of any size) we draw random nested chains $\{c\} \cup \{u_1, \dots, u_k\}$, $k = 1, \dots, \min(K_{\max}, |e|) - 1$, over a random ordering of $e \setminus \{c\}$, and add the resulting subsets as sub tokens (up to a per-hyperedge cap). A hyperedge with $|e| > K_{\max}$ is thus represented only through its sub tokens. Both obs and sub tokens have existence input 1 and form the positive pool P_o of each order o.

Negative subsets. For each positive token e (obs hyperedge or sub token) with $c \in e$, three guarded operations propose negative candidates:

$$\begin{aligned}
\textsc{Drop} &: \text{if } |e| > 2, \text{ sample } v \in e \setminus \{c\}, \text{ emit } e \setminus \{v\}, \\
\textsc{Add} &: \text{if } |e| < K_{\max}, \text{ sample } u \in \mathcal{V} \setminus e, \text{ emit } e \cup \{u\}, \\
\textsc{Swap} &: \text{repeat } r \text{ times: sample } (v, u) \in (e \setminus \{c\}) \times (\mathcal{V} \setminus e), \text{ emit } (e \setminus \{v\}) \cup \{u\},
\end{aligned}$$

with r set per dataset. A candidate S is accepted iff $c \in S$, $2 \le |S| \le K_{\max}$, and S is neither a visible hyperedge nor a sub token of the same view; a subset of an observed hyperedge that was not drawn as a sub token can therefore enter as a neg token. Drop and Add candidates are adjacent in the inclusion lattice to their source, which is what activates the Emer and Inhib branches of the composition label (Section 3); Swap candidates stay at the same order. A candidate becomes the endpoint of a DAG edge only if an adjacent token is retained in the same view.

Bounded sampling and stochastic views. With per-order cap K_o and negative quota B_o^{neg}, each order $o \in \{2, \dots, K_{\max}\}$ is filled observed-heavy: take $n_+ = \min(|P_o|, K_o)$ uniform draws (without replacement) from the positive pool P_o, then $n_- = \min(B_o^{\mathrm{neg}}, K_o - n_+)$ from $N_o = \{S \in \mathcal{N}_c : |S| = o\}$; an order whose positive pool fills the cap therefore receives no neg tokens. The edge set is then recomputed as $\mathcal{R}_c = \{(S, S') \in \mathcal{S}_c^2 : S \subset S',\ |S'| = |S| + 1\}$. We sample V (two for all benchmarks except Walmart, which uses one) independent DAGs per center and concatenate them into a single token sequence of length at most $V(1 + \sum_o K_o)$; each view ends with its own center token. Adjacency-keyed biases (b^{dir}, b^{sib}) are computed view-locally; content-keyed biases (b^{comp}, b^{ord}, b^{ovlp}) fire across views, so the encoder averages compositionality cues over the V samples of $\mathcal{N}_c$.

C.2 Encoder Architecture and Pretraining Heads

Token embedding. Extending the four-attribute lookup of Section 4.2, we add a fifth additive term, the view embedding $\mathbf{e}^{\mathrm{view}}_{v(S)}$, so identical subsets across views remain distinguishable. The feature projector is a 2-layer MLP (Linear, ReLU, Dropout, Linear) at width d; every attribute table reserves index 0 for padding, and the position table has capacity 1024.

Structural biases. The composition bias b_{ij}^{comp} is a 7-row per-head embedding table, zero-initialized, indexed by the asymmetric role pair $(\tau(S_i), \tau(S_j))$ with i the query (Table 4). The asymmetry exploits that every non-center token contains $\{c\}$, so a center query reads informative role signals (buckets 5, 6); mirror cases collapse into bucket 0. The auxiliary term $b^{\text{aux}} = b^{\text{ord}} + b^{\text{ovlp}} + b^{\text{sib}}$ is also per-head and zero-initialized: b^{ord} uses 7 buckets on the signed gap $|S_i| - |S_j|$ at $\{\leq -3, -2, -1, 0, 1, 2, \geq 3\}$; b^{ovlp} uses 5 buckets on the Jaccard ratio $|S_i \cap S_j|/|S_i \cup S_j|$ at thresholds $\{0, 0.25, 0.5, 0.75\}$; b^{sib} is a scalar gated by whether $S_i \neq S_j$ share an immediate parent or child in $\mathcal{G}_c$.

Table 4: Asymmetric bucket index for b_{ij}^{comp}. i is the query, j the key. Bucket 0 is the catch-all for any role tuple outside the six explicit cases (including pairs whose key is the center).

$(\tau(S_i), \tau(S_j))$	Bucket
(obs, obs)	1
(obs, neg)	2
(neg, obs)	3
(neg, neg)	4
(center, obs)	5
(center, neg)	6
all other role tuples	0

Pretraining heads. The semantic head $\text{Pred} : \mathbb{R}^d \to \mathbb{R}^d$ is a 2-layer MLP (Linear, GELU, Dropout, Linear) at width d, regressed with mean squared error against the EMA teacher (a copy of the feature projection, embeddings, Transformer blocks, final LayerNorm, and Pred, updated by EMA with momentum 0.996 and applied to the unmasked sequence; it uses the student's bias tables) on masked positions $\mathcal{M}$. The existence head is a $d \to d/2 \to 1$ bottleneck with GELU and Dropout, emitting the BCE logit at every non-padding token; the source head has the same shape with one output per source tag. The mask token $\mathbf{e}_{\text{mask}} \in \mathbb{R}^d$ is initialized elementwise from $\mathcal{N}(0, 0.02^2)$; the per-batch mask ratio is $\rho = 0.2$ over non-center tokens.

C.3 Finetuning: Weight Transfer and Downstream Readouts

Weight transfer. At finetune, a non-strict state-dict load keeps the encoder backbone and structural-bias tables: `feature_projection`, the five attribute embeddings (`order`, `exist`, `exist_source`, `view`, `position`), the four bias tables (`edge_direction`, `order_distance`, `overlap_bucket`, `exist_transition`), the sibling-bias scalar, every Transformer block, and the final LayerNorm. The pretraining heads, $\mathbf{e}_{\text{mask}}$, the teacher, and the readouts below are reinitialized; Table 1 (HGPM without pretraining) reports the pretrain-vs-scratch comparison.

Node classification. An attention pool $\alpha_S = \text{softmax}_S(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{h}_S^{(L)}))$ is concatenated with the center, $\mathbf{h}_c^{\text{node}} = [\mathbf{h}_{\{c\}}^{(L)} \,\|\, \sum_S \alpha_S \mathbf{h}_S^{(L)}]$, where $\mathbf{h}_{\{c\}}^{(L)}$ averages the V center tokens, and decoded by an MLP under cross-entropy.

Higher-order drug interaction. For a combination $G = \{d_1, \ldots, d_m\}$ conditioned on a side-effect s, we build one inclusion DAG per drug, average their center embeddings $\mathbf{h}_G = \frac{1}{m}\sum_j \mathbf{h}_{\{d_j\}}^{(L)}$, and decode $\hat{y}_{G,s} = \sigma(\text{MLP}([\mathbf{h}_G \,\|\, \mathbf{e}_s]))$ with a learnable $\mathbf{e}_s$ under binary cross-entropy (class-balanced by the positive/negative ratio of the training records). The per-drug DAG of d_j is instantiated on the queried set: its tokens are the subsets of G that contain d_j (up to order 4 on HODDI and 5 on JADER) together with G itself. A token's existence and composition inputs are read from the *training* records with the same side effect s (positive, negative, or unknown when (S, s) is not a training record), and the token for G itself has its existence and composition inputs masked, so no validation or test label enters the input. Each drug is represented by a projection of its SMILES-based feature vector (all-zero when no SMILES is available) plus a learnable drug-identity embedding. The drug encoder is pretrained on node-centered DAGs of HODDI drugs with the objective of Eq. (4) ($\lambda_{\text{exist}} = 0.05$ for HODDI; semantic reconstruction only for JADER).

D DATASETS

D.1 GENERAL HYPERGRAPH BENCHMARKS

We adopt the standard 8-benchmark suite for hypergraph node classification established by AllSet (Chien et al., 2022) and extended with the four heterophilic datasets in ED-HNN (Wang et al., 2023a). The suite splits into four homophilic citation / co-authorship datasets and four structurally diverse / heterophilic datasets. Across all eight datasets we use the same $50/25/25$ random train/validation/test split, and report mean $\pm$ std over ten random seeds. Table 5 summarizes per-dataset statistics; per-dataset descriptions follow.

Table 5: Statistics of the eight hypergraph node classification benchmarks. avg. $|e|$ is the mean hyperedge size; CE-hom. is the clique-expansion homophily ratio (values ≈ 0.5 indicate heterophilic behavior); feat. dim is the dimension of node input features (synthetic CSBM-style Gaussian features, dim 100, are imputed for the heterophilic datasets in the absence of native node attributes).

Dataset	$\lvert V\rvert$	$\lvert E\rvert$	avg. $\lvert e\rvert$	#cls	feat. dim	CE-hom.
Citeseer	3,312	1,079	3.20	6	3703	0.893
Pubmed	19,717	7,963	4.35	3	500	0.952
Cora-CA	2,708	1,072	4.28	7	1433	0.803
DBLP-CA	41,302	22,363	4.45	6	1425	0.869
Congress	1,718	83,105	8.66	2	100	0.555
Senate	282	315	17.17	2	100	0.498
Walmart	88,860	69,906	6.59	11	100	0.530
House	1,290	340	34.73	2	100	0.509

Citeseer (cocitation). Hyperedges represent sets of papers co-cited by another paper. Adapted to the hypergraph format in HyperGCN (Yadati et al., 2019). Node features are bag-of-words abstracts (dim 3703); labels are six paper subject categories.

Pubmed (cocitation). Same cocitation construction as Citeseer, sourced via HyperGCN. Features: dim 500 TF-IDF. Labels: three diabetes subtypes.

Cora-CA (co-authorship). Hyperedges represent sets of papers sharing one author, processed by HyperGCN (Yadati et al., 2019). Features: dim 1433 bag-of-words; labels: seven paper topics.

DBLP-CA (co-authorship). Co-authorship hyperedges from the DBLP citation network, released via HyperGCN (Yadati et al., 2019); the largest co-authorship benchmark in the suite. Features: dim 1425 bag-of-words; labels: six venue/topic classes.

Congress. Bill-cosponsorship network of US Congress members, introduced into the hypergraph-NN literature by ED-HNN (Wang et al., 2023a): nodes are Congresspersons, hyperedges are the sets of (sponsor + co-sponsors) of a bill. With no native attributes, features are imputed as a label-dependent CSBM-style Gaussian (dim 100, $\sigma = 1$; full specification in the Preprocessing paragraph below). Labels: political party (2 classes). Notably dense: average node degree ≈ 427.

Senate. Same construction as Congress, restricted to Senate-only bills. Features: synthetic Gaussian (dim 100, $\sigma = 1$). Labels: party (2 classes). Smallest dataset in the suite ($|V| = 282$); results have unusually high seed-to-seed variance.

Walmart. Customer trip data introduced to the hypergraph-NN literature by AllSet (Chien et al., 2022): nodes are products, hyperedges are sets of products purchased together in a trip. Features: synthetic Gaussian (dim 100, $\sigma = 1$). Labels: 11 product categories. Largest benchmark in the suite ($|V| = 88{,}860$).

House. US House of Representatives committee membership network packaged for hypergraph-NN by AllSet (Chien et al., 2022): nodes are representatives, hyperedges are committee memberships. Features: synthetic Gaussian (dim 100, $\sigma = 1$). Labels: party (2 classes). Notable for very large hyperedges (avg. $|e| \approx 34.7$).

D.2 Drug Interaction Datasets

We evaluate HGPM on two higher-order drug interaction datasets, HODDI (Wang et al., 2025c) and the Japanese Adverse Drug Event Report (JADER) database (Tsuchiya et al., 2020)[1]. Both are processed into the same record format—each record pairs an observed drug combination with a side-effect class and a binary label (positive or negative record)—and use a $50/25/25$ random train/validation/test split. Per-dataset statistics after preprocessing are summarized in Table 6.

Table 6: Statistics of the two drug interaction datasets after preprocessing. The lower block shows the per-combination-size sample count.

Field	HODDI	JADER
#drugs (nodes)	1,821	4,230
#records (samples)	72,986	126,334
#side-effect classes	1,710	100 (filtered)
Combo size range	2–8	3–8
Mean combo size (all records)	2.61	3.84
Positive / Negative	42,113 / 30,873	63,167 / 63,167
Train / Val / Test	36,493 / 18,246 / 18,247	63,116 / 31,508 / 31,710
Size 2	49,172	—
Size 3	12,764	68,722
Size 4	5,382	30,914
Size 5	3,252	13,960
Size 6	1,278	6,422
Size 7	689	4,040
Size 8	449	2,276

HODDI. HODDI (Wang et al., 2025c) is a higher-order drug–side-effect benchmark derived from FAERS spontaneous-reporting case reports. After deduplication, drug-name normalization, and side-effect filtering, each combination of 2–8 drugs is associated with one of 1,710 side-effect classes. We download HODDI from the authors' release at `https://github.com/TIML-Group/HODDI` (MIT license) and apply our $50/25/25$ random split protocol; this differs from the quarter-based split used in the original HODDI paper.

JADER. JADER (Japanese Adverse Drug Event Report) is the spontaneous-reporting pharmacovigilance database maintained by Japan's Pharmaceuticals and Medical Devices Agency (PMDA), publicly available at `https://www.info.pmda.go.jp/fukusayoudb/CsvDownload`. Each case report is split across four relational tables (DEMO, DRUG, REAC, HIST) joined by a case identifier. We construct hyperedges from the DRUG table by grouping drugs reported in the same case (sizes 3–8 retained) and label each hyperedge with one of 100 retained MedDRA-coded side-effect classes from the REAC table, following a processing pipeline analogous to HODDI's so that the two datasets share a common hyperedge-and-label format (Table 6). Negatives are sampled to match positives at a 1:1 ratio. Concretely, we use the April 2026 release (`pmdacasereport202604`), keep cases with 3–8 suspect drugs (only the suspect-drug role) and at least one reported adverse event, retain adverse-event classes with 100–1,500 positive (drug set, adverse event) pairs and keep the 100 most frequent of them, sample each negative from a case with the same number of drugs that does not report the target adverse event, and split records $50/25/25$ stratified by adverse event and label. JADER is provided by PMDA under terms of use that do not permit redistribution; results obtained with it are reported as such, and we release only the preprocessing scripts.

[1] `https://www.info.pmda.go.jp/fukusayoudb/CsvDownload`

Tasks. On both datasets we evaluate two predictive tasks, both binary classifications of (drug set, side effect) records with the readout of Appendix C.3: *edge classification* (whether an observed drug combination is reported with a given side effect), reported with F1 at threshold 0.5 and AUROC; and *link prediction* (whether a combination–side-effect record is observed rather than a negative record), reported with AUROC and AUPRC (average precision). Cross-order generalization (predict higher-order interactions from lower-order observed context) is evaluated only on HODDI (Section 6).

E Implementation Details

E.1 Hyperparameters

Pretraining. The pretrain stage shares the encoder hyperparameters $(d, L, H, \text{dropout})$ with the downstream model selected for the same benchmark, so that the finetune step can load the pretrained encoder without reshaping. Pretraining also uses the benchmark's optimization settings (epochs, batch size, learning rate, weight decay, gradient clip, patience). The remaining pretraining values are shared: mask ratio $\rho = 0.2$ over non-center tokens (at least one per sequence), early stopping on the pretraining loss of a random held-out 10% of the target DAGs, EMA momentum 0.996, AdamW, and $\lambda_{\text{exist}} = \lambda_{\text{src}} = 0$ except on Congress (1). The pretraining corpus contains the DAGs of all nodes, including validation and test nodes, but no labels (transductive setting).

Downstream search. For each benchmark we run a per-dataset random search over the discrete grid in Table 7. The candidate with the highest mean validation accuracy is selected. The optimizer is AdamW throughout. The maximum number of training epochs is held fixed per dataset; the actual stopping point is determined by early stopping on validation accuracy with the patience drawn from the table.

Table 7: Hyperparameter search space for HGPM downstream training. Ranges are the union of values explored across the eight benchmarks via per-dataset random search. Some values used for the reported results (e.g., $K_o = 24$, $r = 8$, and Walmart's $K_{\max} = 4$, $K_o = 2$) lie outside this grid.

Hyperparameter	Search range
Encoder.	
Hidden dimension d	$\{32, 64, 128, 256, 512, 768\}$
Layers L	$\{2, 3, 4, 5, 6\}$
Attention heads H	$\{2, 4, 8, 16\}$ (must divide d)
Dropout	$\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$
Inclusion DAG.	
Max order $K_{\max}$	$\{3, 5, 8, 10, 12, 16, 18, 24\}$
Per-order budget K_o	$\{4, 6, 8, 10, 12, 16, 20\}$
SWAP proposals per positive r	$\{1, 2, 3\}$
Stochastic views V	2 (1 for Walmart)
Optimization.	
Learning rate	$\{3{\cdot}10^{-5}, 10^{-4}, 3{\cdot}10^{-4}, 5{\cdot}10^{-4}, 10^{-3}, 3{\cdot}10^{-3}, 10^{-2}\}$
Weight decay	$\{0, 10^{-5}, 10^{-4}, 5{\cdot}10^{-4}, 10^{-3}, 5{\cdot}10^{-3}, 10^{-2}, 5{\cdot}10^{-2}\}$
Batch size	$\{16, 32, 64, 128, 256, 512, 1024\}$
Gradient clip $\|g\|_2$	$\{0.25, 0.5, 1.0, 2.0, 5.0\}$
Early-stop patience	$\{3, 5, 8, 10, 12, 15, 20, 50, 100\}$

E.2 Compute Resources

All experiments—pretraining, finetuning, and ablations—are conducted on NVIDIA A40 GPUs (48 GB). Each training run uses a single GPU. The host machine uses dual Intel Xeon Gold 6326 CPUs (2.90 GHz, 16 cores each) running Red Hat Enterprise Linux 9.7 with Linux kernel 5.14.

Table 8: Full per-dataset results for the six HGPM model-space ablations. Within each ablation block, the gray and **bold** row is the HGPM default; the default attains the best result in every column. Numbers are mean test accuracy over 10 seeds.

	Homophilic				Heterophilic			
Setting	Citeseer	Pubmed	Cora-CA	DBLP-CA	Congress	Senate	Walmart	House
(a) Subset Tokens								
Observed hyperedges only	$73.27_{\pm 1.6}$	$88.04_{\pm 0.7}$	$82.43_{\pm 1.7}$	$89.97_{\pm 0.4}$	$89.73_{\pm 1.3}$	$71.55_{\pm 5.2}$	$65.41_{\pm 0.8}$	$75.13_{\pm 2.3}$
Uncentered random subsets	$75.42_{\pm 1.4}$	$88.86_{\pm 0.6}$	$83.78_{\pm 1.4}$	$91.07_{\pm 0.3}$	$91.46_{\pm 1.1}$	$74.58_{\pm 4.6}$	$67.43_{\pm 0.6}$	$76.58_{\pm 1.9}$
No perturbation negatives	$76.78_{\pm 1.2}$	$89.51_{\pm 0.5}$	$84.93_{\pm 0.6}$	$92.04_{\pm 0.2}$	$93.42_{\pm 0.9}$	$76.34_{\pm 5.0}$	$70.07_{\pm 0.4}$	$78.61_{\pm 1.6}$
Inclusion DAG tokens	$\mathbf{77.10_{\pm 1.2}}$	$\mathbf{89.70_{\pm 0.5}}$	$\mathbf{85.10_{\pm 0.2}}$	$\mathbf{92.20_{\pm 0.2}}$	$\mathbf{94.10_{\pm 1.0}}$	$\mathbf{77.20_{\pm 4.8}}$	$\mathbf{71.20_{\pm 0.5}}$	$\mathbf{79.20_{\pm 1.8}}$
(b) Inclusion DAG Edges								
No edges (flat tokens)	$74.13_{\pm 1.8}$	$87.83_{\pm 0.8}$	$82.61_{\pm 1.8}$	$90.07_{\pm 0.5}$	$89.81_{\pm 1.4}$	$74.59_{\pm 5.0}$	$66.41_{\pm 0.9}$	$75.83_{\pm 2.4}$
Random edges	$75.34_{\pm 1.5}$	$88.46_{\pm 0.7}$	$83.71_{\pm 1.4}$	$90.83_{\pm 0.4}$	$91.69_{\pm 1.2}$	$75.68_{\pm 4.9}$	$68.18_{\pm 0.7}$	$77.04_{\pm 2.1}$
Center-only edges	$76.21_{\pm 1.3}$	$89.07_{\pm 0.5}$	$84.36_{\pm 1.0}$	$91.62_{\pm 0.3}$	$92.72_{\pm 1.0}$	$76.51_{\pm 4.8}$	$69.62_{\pm 0.6}$	$78.07_{\pm 1.7}$
Adjacent-order inclusion edges	$\mathbf{77.10_{\pm 1.2}}$	$\mathbf{89.70_{\pm 0.5}}$	$\mathbf{85.10_{\pm 0.2}}$	$\mathbf{92.20_{\pm 0.2}}$	$\mathbf{94.10_{\pm 1.0}}$	$\mathbf{77.20_{\pm 4.8}}$	$\mathbf{71.20_{\pm 0.5}}$	$\mathbf{79.20_{\pm 1.8}}$
(c) Subset-Token Encoder								
Mean pooling	$73.92_{\pm 1.6}$	$87.62_{\pm 0.8}$	$81.83_{\pm 1.4}$	$89.78_{\pm 0.5}$	$89.43_{\pm 1.3}$	$74.37_{\pm 5.4}$	$65.93_{\pm 0.8}$	$75.42_{\pm 2.2}$
GRU over linearized sequence	$71.43_{\pm 3.3}$	$84.31_{\pm 2.7}$	$77.84_{\pm 3.9}$	$84.62_{\pm 3.2}$	$86.71_{\pm 2.8}$	$67.89_{\pm 6.3}$	$64.13_{\pm 2.5}$	$72.93_{\pm 3.4}$
Transformer w/o structural biases	$75.61_{\pm 1.5}$	$88.71_{\pm 0.6}$	$83.74_{\pm 1.4}$	$91.06_{\pm 0.4}$	$91.32_{\pm 1.2}$	$75.68_{\pm 4.9}$	$67.83_{\pm 0.7}$	$77.13_{\pm 2.0}$
Inclusion-aware Transformer	$\mathbf{77.10_{\pm 1.2}}$	$\mathbf{89.70_{\pm 0.5}}$	$\mathbf{85.10_{\pm 0.2}}$	$\mathbf{92.20_{\pm 0.2}}$	$\mathbf{94.10_{\pm 1.0}}$	$\mathbf{77.20_{\pm 4.8}}$	$\mathbf{71.20_{\pm 0.5}}$	$\mathbf{79.20_{\pm 1.8}}$
(d) Structural Attention Biases								
No bias (vanilla attention)	$75.61_{\pm 1.5}$	$88.71_{\pm 0.6}$	$83.74_{\pm 1.4}$	$91.06_{\pm 0.4}$	$91.32_{\pm 1.2}$	$75.68_{\pm 4.9}$	$67.83_{\pm 0.7}$	$77.13_{\pm 2.0}$
Composition only	$76.72_{\pm 1.3}$	$89.46_{\pm 0.5}$	$84.78_{\pm 0.8}$	$91.89_{\pm 0.3}$	$93.47_{\pm 0.9}$	$76.81_{\pm 5.1}$	$70.53_{\pm 0.6}$	$78.58_{\pm 1.7}$
Direction + auxiliary	$76.13_{\pm 1.4}$	$89.18_{\pm 0.6}$	$84.31_{\pm 1.1}$	$91.56_{\pm 0.3}$	$92.68_{\pm 1.1}$	$76.35_{\pm 5.2}$	$69.43_{\pm 0.6}$	$77.96_{\pm 1.9}$
Full bias	$\mathbf{77.10_{\pm 1.2}}$	$\mathbf{89.70_{\pm 0.5}}$	$\mathbf{85.10_{\pm 0.2}}$	$\mathbf{92.20_{\pm 0.2}}$	$\mathbf{94.10_{\pm 1.0}}$	$\mathbf{77.20_{\pm 4.8}}$	$\mathbf{71.20_{\pm 0.5}}$	$\mathbf{79.20_{\pm 1.8}}$
(e) Pretraining Reconstruction Target								
Existence label only	$73.35_{\pm 2.5}$	$87.28_{\pm 1.3}$	$81.88_{\pm 2.4}$	$90.83_{\pm 1.1}$	$92.23_{\pm 1.8}$	$74.56_{\pm 5.1}$	$68.44_{\pm 1.1}$	$77.38_{\pm 2.6}$
Raw subset feature	$75.15_{\pm 1.8}$	$88.38_{\pm 0.9}$	$83.14_{\pm 1.8}$	$91.38_{\pm 0.7}$	$93.00_{\pm 1.3}$	$75.48_{\pm 4.9}$	$69.82_{\pm 0.8}$	$78.22_{\pm 2.2}$
Composition label only	$75.90_{\pm 1.4}$	$89.04_{\pm 0.6}$	$84.12_{\pm 1.3}$	$91.65_{\pm 0.5}$	$93.59_{\pm 1.1}$	$76.33_{\pm 4.8}$	$70.40_{\pm 0.6}$	$78.78_{\pm 1.8}$
EMA-teacher target	$\mathbf{77.10_{\pm 1.2}}$	$\mathbf{89.70_{\pm 0.5}}$	$\mathbf{85.10_{\pm 0.2}}$	$\mathbf{92.20_{\pm 0.2}}$	$\mathbf{94.10_{\pm 1.0}}$	$\mathbf{77.20_{\pm 4.8}}$	$\mathbf{71.20_{\pm 0.5}}$	$\mathbf{79.20_{\pm 1.8}}$
(f) Token Sequence Order								
Ascending order, center first	$76.11_{\pm 1.8}$	$89.18_{\pm 0.8}$	$84.31_{\pm 1.4}$	$91.63_{\pm 0.6}$	$92.92_{\pm 1.4}$	$75.46_{\pm 5.4}$	$69.73_{\pm 1.0}$	$78.21_{\pm 2.4}$
Shuffled within order	$76.57_{\pm 1.5}$	$89.43_{\pm 0.7}$	$84.67_{\pm 1.1}$	$91.88_{\pm 0.4}$	$93.47_{\pm 1.2}$	$76.12_{\pm 5.1}$	$70.41_{\pm 0.8}$	$78.68_{\pm 2.0}$
Fully shuffled (non-topological)	$76.93_{\pm 1.3}$	$89.61_{\pm 0.6}$	$84.93_{\pm 0.8}$	$92.08_{\pm 0.3}$	$93.89_{\pm 1.1}$	$76.81_{\pm 5.0}$	$70.94_{\pm 0.6}$	$79.01_{\pm 1.8}$
Descending-order linearization	$\mathbf{77.10_{\pm 1.2}}$	$\mathbf{89.70_{\pm 0.5}}$	$\mathbf{85.10_{\pm 0.2}}$	$\mathbf{92.20_{\pm 0.2}}$	$\mathbf{94.10_{\pm 1.0}}$	$\mathbf{77.20_{\pm 4.8}}$	$\mathbf{71.20_{\pm 0.5}}$	$\mathbf{79.20_{\pm 1.8}}$

E.3 Software and Library Versions

We use Python 3.11 and PyTorch 2.6.0 with the CUDA 12.4 build (cuDNN 9.1). Other key dependencies: NumPy 2.4, SciPy 1.17, scikit-learn 1.8, PyYAML 6.0, and Weights & Biases 0.26.

E.4 Configurations Used for the Reported Results

Pretraining and fine-tuning share the model and optimization settings. For the drug tasks, the encoder uses $d = 256$, $L = 4$, $H = 8$, and dropout 0.1; fine-tuning uses 10 epochs, batch size 48, learning rate $2 \cdot 10^{-4}$, weight decay 10^{-2}, and patience 3 on HODDI, and 16 epochs, batch size 64, learning rate $3 \cdot 10^{-4}$, weight decay 10^{-3}, and patience 4 on JADER, with a linear warm-up and cosine schedule; the model is selected by validation AUROC (HODDI) or AUPRC (JADER).

F Comprehensive Ablation Study

The six ablations summarized in §5 (Table 2) report only the homophilic and heterophilic group averages for compactness. Table 8 below gives the full per-dataset breakdown over all eight benchmarks (Citeseer, Pubmed, Cora-CA, DBLP-CA, Congress, Senate, Walmart, House) for all six ablations. Within each ablation block, the gray cells (also **bold**) mark the HGPM default, which is best in every column. Numbers are mean test accuracy over 10 seeds.

(a) Subset Tokens. The four rows isolate three independent design choices via three pairwise comparisons. Substructure expansion (Observed hyperedges only vs. No perturbation negatives, both centered) contributes the largest average gain among the three comparisons, with the gap

concentrated on heterophilic and sparse-hyperedge data and shrinking on saturated co-citation data (see Table 8). The mechanism is that subset tokens expose the COMP/EMER/INHIB structure defined in Section 3, which a representation built only from hyperedges considered in isolation does not carry. Center anchoring (Uncentered random subsets vs. No perturbation negatives) contributes a smaller but consistent gain across datasets by concentrating the fixed token budget on subsets containing the target node c. Sampled negatives (No perturbation negatives vs. Default) yield the smallest contribution on average and are dataset-dependent, falling within noise on saturated benchmarks while providing a modest gain on heterophilic ones where INHIB regimes carry real signal. Substructure expansion thus accounts for most of the gain, with negative-token regularization a secondary effect.

(b) Inclusion DAG Edges. The four rows isolate three properties of the structural attention bias. The total contribution of the structural bias (Default vs. No-edges) is $+3.07$pp on average, with the largest gaps on dense heterophilic data and the smallest on saturated co-citation data. Random edges recover a minority of this gain while center-only edges recover the majority, indicating that the *content* of inclusion edges, namely adjacent-order subset-to-superset adjacency, matters far more than their quantity. Center anchoring is therefore the dominant structural prior, with adjacent-order topology providing the remaining signal that uniquely encodes the COMP/EMER/INHIB regimes (since composition labels are defined only on adjacent-order pairs). **Cross-table tie-in.** Tables (a) and (b) independently identify center anchoring as the dominant factor: (a) at the tokenization layer (subsets must contain c) and (b) at the attention-topology layer (every token attends preferentially to c).

(c) Subset-Token Encoder. The ranking GRU $<$ mean pooling $<$ Transformer without structural biases $<$ inclusion-aware Transformer (default) holds across datasets, with each step contributing a different mechanism. The GRU consistently underperforms mean pooling, which is consistent with the inclusion-DAG token sequence having no canonical order so that imposing a sequential order is uninformative; the effect is largest on small-data Senate, where the absolute GRU gap to default is the widest. Mean pooling is a strong baseline that recovers roughly half of the Transformer's improvement via feature averaging alone, and the vanilla Transformer adds attention contextualization on top. The structural bias contributes the remaining gap (Default vs. Vanilla), and is largest on heterophilic dense data and smallest on saturated co-citation data. **Cross-table tie-in.** The (c) Transformer-w/o-biases row is identical to the (d) No-bias row. The (b) No-edges (flat) row sits below (c) Transformer-w/o-biases because (b) additionally removes the topological position embedding; this residual gap is several times larger than the (f) shuffled-vs-default gap, consistent with position embedding becoming the primary carrier of topological signal once the structural bias is disabled.

(d) Structural Attention Biases. This decomposes the structural attention bias of Eq. (3) into its semantic (b^{comp}) and topological ($b^{\mathrm{dir}} + b^{\mathrm{aux}}$) components. The total bias gap (Default vs. No-bias) averages about 1.84pp. The role-pair (composition) bias alone recovers 70–80% of this gap on each benchmark, roughly three quarters on average, with the recovery stable across benchmarks regardless of homophily, saturation, or scale. Direction $+$ auxiliary biases together recover a smaller fraction on average; the two recoveries are not additive, which is consistent with partial redundancy between composition and topology biases. This supports the design choice in Eq. (3) of including b^{comp} as an explicit attention bias rather than relying on topology priors alone.

(e) Pretraining Reconstruction Target. Unlike the previous rows, which ablate architecture, (e) ablates the pretraining target while keeping the architecture fixed. The four targets form a clear hierarchy on every benchmark: existence-only (predicting $\mathbb{1}_S$) underperforms even training from scratch, raw subset features give modest gains, composition labels (discrete COMP/EMER/INHIB supervision) recover the bulk of the gap, and the EMA-teacher target attains the default. Existence-only supplies too narrow a signal: the encoder fits a binary indicator that does not transfer to downstream node classification. Raw subset features are trivially recoverable from mean-pooling, which limits their pretraining utility. Composition-label reconstruction is closer to the default, and the continuous teacher target performs best; adjacent steps are within seed variance on the small benchmarks.

(f) Token Sequence Order. This row probes sensitivity to the token order fed to the Transformer. All linearization variants stay close to default, with sub-1pp gaps from the fully shuffled setting on

average, indicating that structural attention bias rather than position embedding carries the topological signal. The descending-order linearization remains the strongest by a small margin, and the model is largely order-robust as long as the bias terms remain active. **Cross-table tie-in.** The shuffled-vs-default gap in (f) is much smaller than the No-edges to Transformer-without-biases gap spanning (b) and (c). Both comparisons contrast topology-via-position with no-position, but in (f) the structural bias is *active* and carries the topological signal redundantly, whereas in the (b) to (c) comparison the bias is *off* and position embedding is the only carrier, producing a substantially larger gap.

Cross-axis trends: heterophilic data benefits more from architectural priors. A consistent pattern emerges across the four architectural ablations from (a) through (d): each non-catastrophic design choice has a larger impact on the heterophilic group than on the homophilic group, holding for Hyperedge-only in (a), No-edges in (b), Vanilla Transformer in (c), and No-bias in (d). The catastrophic GRU row of (c) is the exception, since both groups collapse to a similar absolute degree when the inductive bias is fundamentally wrong rather than merely missing. We note that heterophilic datasets also have lower base accuracy and therefore more headroom, which contributes to the larger absolute gaps. The pretraining-target ablation (e) does *not* show the same asymmetry; existence-only even widens the gap slightly more on the homophilic group, suggesting that representation-quality choices and inductive-bias choices have qualitatively different per-group impact, and that HGPM's heterophilic advantage is more closely tied to architectural priors than to the pretraining target. Linearization (f) follows the architectural pattern at much smaller magnitudes. Because the heterophilic benchmarks use label-dependent synthetic features (Appendix D.1), we do not interpret this asymmetry mechanistically.

G Case Study: HGPM vs. Baseline

We complement the case study in Figure 4 with two head-to-head comparisons against KHGNN (Xie et al., 2025), the strongest baseline in Table 3, on the same three candidates added to FOLFOX (panitumumab, bevacizumab, capecitabine). Both panels share the JADER-derived labels used in the main text: the FOLFOX + panitumumab record is reported with peripheral neuropathy (positive, COMP), whereas the FOLFOX + bevacizumab and FOLFOX + capecitabine records are not (negative, INHIB).

Link-prediction score. Figure 5 contrasts each model's binary score against the recorded label. HGPM separates all three correctly: 0.84 for panitumumab (positive), 0.29 for bevacizumab and 0.22 for capecitabine (both negative, well below the 0.5 threshold). The baseline scores all three above the threshold ($0.58/0.63/0.54$) and therefore flags bevacizumab and capecitabine as positive, in opposition to the data. A possible explanation, which we do not test, is that the baseline's score follows drug-feature similarity (panitumumab and bevacizumab are close in feature space) rather than the joint observation pattern of the 3-drug subset and its 4-drug superset.

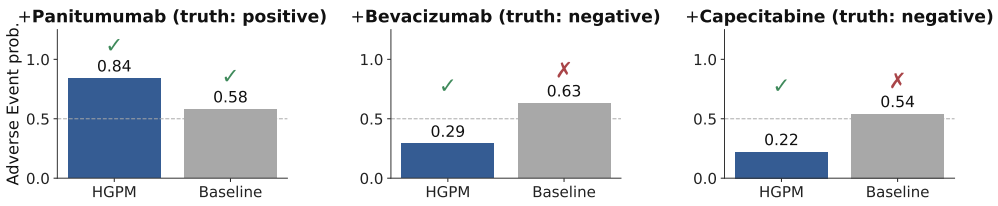


Figure 5: Binary link-prediction score for each candidate added to FOLFOX, comparing HGPM with KHGNN. The recorded label is shown above each card; the dashed line marks the 0.5 decision threshold. HGPM separates all three branches; the baseline misclassifies bevacizumab and capecitabine as positive.

Adverse-event distribution. Figure 6 shows the baseline's top-3 predicted adverse events per candidate. Peripheral neuropathy is ranked first across all three branches, with probabilities $0.44/0.26/0.28$, mirroring the link-prediction failure: the baseline averages all three candidates toward a shared neuropathy prior. HGPM (Figure 4d) demotes peripheral neuropathy to rank 2 for bevacizumab and rank

3 for capecitabine, surfacing regimen-specific signatures instead: hypertension and proteinuria for the anti-VEGF branch, and hand-foot syndrome and diarrhea for the fluoropyrimidine branch. These rankings are consistent with the labeled toxicities of the anti-VEGF and fluoropyrimidine classes; the adverse-event head is trained on JADER adverse-event labels. The comparison is qualitative, concerns reporting patterns in spontaneous reports, and does not establish any causal effect.

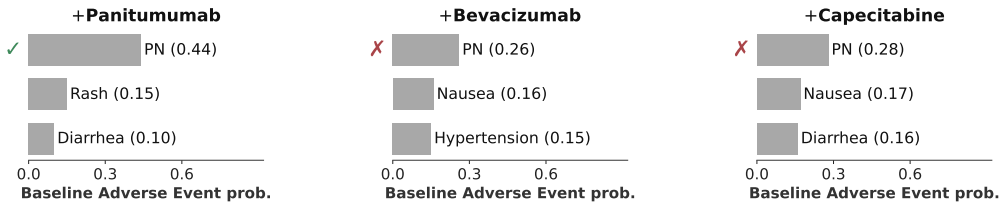


Figure 6: Top-3 adverse events predicted by KHGNN for each candidate added to FOLFOX. Peripheral neuropathy ranks first across all three branches, whereas HGPM ranks candidate-specific adverse events higher.