# INCIDENCE-STRUCTURED SPATIAL FACTORS FOR LATENT CAUSAL DISCOVERY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Latent causal discovery compresses many spatial sites into a few latent time series and learns the causal graph among them. Existing methods link latents to sites through kernels over site coordinates, learned site partitions or principal components. Spatial data often carry memberships such as regions, and the hypergraph lift of a kernel sees them only through their pairwise operator. We show that this operator does not in general determine the latent coordinates that the memberships determine. Two incidences with the same pairwise operator admit models with the same observed law and inequivalent latent laws. We therefore propose incidence-structured spatial factors, which build each latent mode as a weighted union of structural units. For the unsmoothed factor, we prove the latents identifiable up to permutation, scaling and translation under site-specific smooth invertible observation maps when every mode owns an anchor site and bridge sites connect the modes. On a public grid benchmark, the factor recovers the latent causal graph more accurately than kernel, single-parent, two-stage and sequence-model baselines. On random committee hypergraphs with membership-following footprints, hyperedge atoms, given or reconstructed from the pairwise operator, recover the latents more accurately than pair atoms and heat kernels built on that operator.

## 1 INTRODUCTION

Spatiotemporal systems such as climate fields, sensor networks and neural recordings are observed at many more sites than there are mechanisms that drive them (Runge et al., 2015; 2019). Latent causal discovery addresses this regime by compressing the $L$ observed sites into $D \ll L$ latent time series and learning a temporal causal graph among the latents (Tibau et al., 2022; Brouillard et al., 2026). The recovered graph then relates mechanisms rather than individual sites.

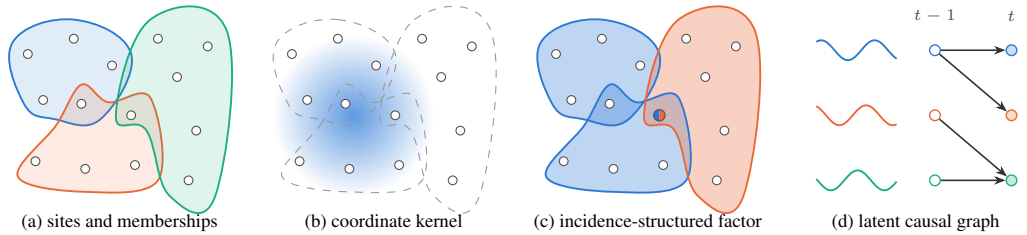
The spatial factor maps latents to sites and carries every spatial assumption. Existing methods differ mainly in how they build it. Two-stage pipelines take Varimax-rotated or tensor principal components and run causal discovery on their scores (Kaiser, 1958; Tibau et al., 2022; Babii et al., 2025). Joint models learn the factor together with the graph. Some use kernels over site coordinates, from topographic factor analysis (Manning et al., 2014) to its causal extension SPACY (Wang et al., 2025). Others learn a single-parent partition of the sites (Brouillard et al., 2026). Sequence models for temporally causal latent processes impose no spatial factor (Yao et al., 2022a;b).

None of these factors takes as input the membership information that spatial data frequently carry: counties belong to states and metropolitan areas, stations lie on transit lines, and gauges share catchments. Each such relation is a hypergraph incidence structure. A coordinate kernel assigns every mode a smooth footprint determined by the site coordinates alone, which crosses membership boundaries (Figure 1b). The hypergraph lift of this kernel is the heat kernel of the normalized hypergraph Laplacian. This heat kernel and spectral hypergraph convolutions built on the same Laplacian act on the memberships only through their weighted clique expansion (Zhou et al., 2006; Agarwal et al., 2006; Chitra & Raphael, 2019; Feng et al., 2019). Two membership relations with the same pairwise operator therefore yield the same factor family.

We show that this reduction loses information about the latent representation itself. We construct two incidences on six sites with the same complete weighted pairwise operator, degrees and unit sizes. Each incidence, taken as fixed, identifies the latent coordinates. Yet the two admit models with the same observed law whose latent laws are not related by any monomial affine map (Theorem 2). The



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.




(a) sites and memberships (b) coordinate kernel (c) incidence-structured factor (d) latent causal graph

Figure 1: (a) Sites (circles) and three memberships. (b) A coordinate kernel gives each latent mode a footprint determined by the site coordinates, which ignores the memberships. (c) An incidence-structured factor builds each mode from whole units; units of the blue mode and of the orange mode meet at a bridge site, which loads on both modes. (d) The latent time series and their temporal causal graph, the object of latent causal discovery.

pairwise operator therefore does not in general determine the latent coordinates that the incidence determines. A spatial factor should read the incidence itself.

We propose *incidence-structured spatial factors* (ISF; Figure 2). Every latent mode is a learned, degree-normalized weighted union of the structural units of the given incidence. Every unit is assigned to at most one mode, so a site covered by units of several modes has several latent parents. An optional heat kernel smooths the footprints. For the unsmoothed family, we prove that the latents are identifiable up to permutation, scaling and translation, provided that every mode owns an anchor site and bridge sites connect the modes (Theorem 1). The result holds under site-specific smooth invertible observation maps, a known noise law and dependent latents. The conditions are read off the incidence and the unit assignment. Our training recipe does not refer to the factor family, so comparisons under a shared recipe isolate the spatial factor.

On a public grid benchmark with the grid adjacency as incidence, the factor recovers the latent causal graph more accurately than kernel, single-parent, two-stage and sequence-model baselines on fresh independent draws. Its advantage over the coordinate kernel persists when sparsity is selected without ground truth. A rewiring control shows that its gain requires the correct incidence. On random committee hypergraphs whose footprints follow the memberships, hyperedge atoms recover the latents better on every structure than pair atoms, a heat kernel on the same pairwise operator and rewired memberships. Hyperedges reconstructed from the weighted clique expansion retain this advantage on almost every structure. On committee structures that satisfy the anchor and bridge conditions, the unsmoothed estimator in the class of the theorem recovers the latents almost exactly on almost every structure. It recovers them accurately on every such structure also when every site is observed through its own monotone map from a known parametric family whose parameters it learns. Our contributions are:

- A separation result for incidence-structured factor models: the pairwise operator of a hypergraph, on which heat kernels and spectral hypergraph convolutions act, does not in general determine the latent coordinates that its incidence determines (Sections 4 and 5.1; Appendix I);
- Incidence-structured spatial factors with identifiability conditions read off the incidence and its unit assignment, and an incidence family that is identifiable without anchors (Sections 4 and 5);
- Experiments in which an estimator improves latent causal discovery on a public grid benchmark over kernel, single-parent, two-stage and sequence-model baselines, and unit-level atoms outperform pair-level atoms and heat kernels on random committee hypergraphs (Section 6).

## 2 Related Work

**Latent causal discovery from spatiotemporal data.** Two-stage pipelines reduce dimension and then run a causal discovery algorithm, as in Mapped-PCMCI (Kaiser, 1958; Runge et al., 2015;

2019; Tibau et al., 2022) or tensor principal components (Babii et al., 2025) followed by PCMCI+ (Runge, 2020). Joint methods learn the reduction and the graph in one objective. Wang et al. (2025) place a radial basis spatial factor over site coordinates inside a variational model whose latent structural causal model follows Gong et al. (2023), with an acyclicity penalty in the style of Zheng et al. (2018). Single-parent decoding assigns every observed variable one latent parent (Brouillard et al., 2026), and sequence models identify temporally causal latent processes without a spatial factor (Yao et al., 2022a;b). We replace the spatial factor and keep a variational formulation in which the factor, the encoder and the graph are trained together. On the grid benchmark, the graph is then relearned on the frozen representation (Section 6.2).

**Known memberships in factor models.** Loadings tied to membership sets appear outside causal discovery: PLIER aligns loadings with sparse combinations of known gene-set memberships (Mao et al., 2019), and overlapping-cluster factor models estimate unknown memberships and identify loadings through pure variables (Bing et al., 2020). Our factor differs in having far fewer modes than units, a learned unit-to-mode assignment, degree-normalized and optionally heat-smoothed footprints, and latents that follow a temporal structural causal model.

**Identifiability.** Post-nonlinear mixtures are classically identified through source independence (Achard & Jutten, 2005). Lyu & Fu (2022) remove the nonlinearity of an overdetermined post-nonlinear mixture without independence for generic mixing matrices, and Theorem 3 gives an affine reduction of this kind for indecomposable factors. Identifiability under other restrictions follows from auxiliary variables (Khemakhem et al., 2020), structural sparsity (Zheng et al., 2022), grouped observations (Morioka & Hyvärinen, 2024), anchor features in deep generative models (Moran et al., 2022) and support conditions for sparse factor analysis (Sturma et al., 2026). Separable nonnegative factorizations and anchor-word topic models are identified or efficiently recovered when, among other conditions, every factor owns a row that loads on it alone (Donoho & Stodden, 2003; Arora et al., 2012; 2013), and pure-measurement conditions identify linear latent causal structure (Silva et al., 2006; Cai et al., 2019; Xie et al., 2020). Condition (C1) of Theorem 1 is of this kind. Theorem 1 reads anchors and bridges off a given incidence and its unit assignment, and combines them with site-specific nonlinear maps and dependent latents through the affine reduction. The rank test behind our anchor-free family is that of Li et al. (2017), which we certify for one incidence family.

**Hypergraphs and their pairwise projections.** The normalized hypergraph Laplacian of Zhou et al. (2006) reduces to a Laplacian of a pairwise expansion (Agarwal et al., 2006; Chitra & Raphael, 2019), and spectral hypergraph convolutions filter with the same operator (Feng et al., 2019). Total-variation formulations instead retain higher-order information (Hein et al., 2013). Distinct hypergraphs can share one pairwise projection (Kim & Kirkland, 2022; Wang & Kleinberg, 2024). Theorem 2 lifts this non-uniqueness from the incidence to the latent coordinates.

## 3 Problem setup

**Incidence structure.** Let $\mathcal{H} = (V, E, w)$ be a connected hypergraph with no isolated site, with $|V| = L$ sites and $|E| = M$ structural units. The incidence matrix is $H \in \{0,1\}^{L\times M}$, with $H_{ve} = 1$ when site $v$ belongs to unit $e$, and $W = \mathrm{diag}(w_e)$ holds the positive unit weights. With weighted site degree $d_v = \sum_e w_e H_{ve}$, we write $\bar{H} = \mathrm{diag}(d)^{-1}HW$ for the degree-normalized incidence.

**Observations and generative model.** We observe $N$ independent sequences of site signals of length $T$, and $X^{(t)} \in \mathbb{R}^L$ is the signal at time $t$ with the sequence index suppressed. The generative model is a factor model with a site-specific observation map,

$$X_v^{(t)} = g_v\Big(F_v^\top Z^{(t)}\Big) + \varepsilon_v^{(t)}, \qquad v \in V,\ t \in [T], \tag{1}$$

where $Z^{(t)} \in \mathbb{R}^D$ collects the $D$ latent time series, $F \in \mathbb{R}^{L\times D}$ is the spatial factor with $v$-th row $F_v^\top$, $g_v : \mathbb{R} \to \mathbb{R}$ is the observation map at site $v$, and $\varepsilon_v^{(t)}$ is observation noise. Column $F_{\cdot d}$ is the *footprint* of latent mode $d$, and the noiseless mixing map is $m_\theta(z) = (g_v(F_v^\top z))_{v\in V}$ with $\theta = (F, g, P_Z)$ and $P_Z$ the law of $Z$.

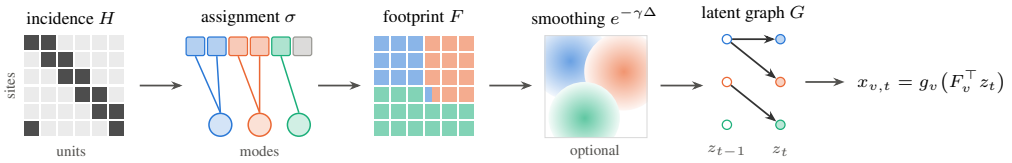



Figure 2: Incidence-structured spatial factor. The incidence $H$ supplies the units, and the assignment $\sigma$ sends each unit to at most one mode (gray: unassigned). In $F = \bar{H}A$, the column of a mode is the degree-normalized union of its units, and a site covered by units of two modes loads on both. An optional heat kernel smooths each footprint. Site $v$ observes its latent parents through $g_v$, and the latents follow a temporal causal graph $G$.

**Latent structural causal model.** The latents follow a temporal structural causal model with maximum lag $\tau$,

$$Z_d^{(t)} = f_d\Big(\mathrm{Pa}_G^d(<t),\, \mathrm{Pa}_G^d(t)\Big) + \eta_d^{(t)}, \qquad G \in \{0,1\}^{(\tau+1)\times D\times D}, \tag{2}$$

where $\mathrm{Pa}_G^d(<t)$ and $\mathrm{Pa}_G^d(t)$ are the lagged and instantaneous parents of $Z_d$ in $G$ and $\eta_d^{(t)}$ is the innovation. $G^{(0)}$ holds instantaneous edges and must be acyclic, and $G^{(1:\tau)}$ holds lagged edges. The grid-benchmark generator uses the functional parameterization and history-dependent spline noise of Gong et al. (2023). Our implementation uses linear or Rhino-inspired nonlinear structural functions with Gaussian innovations, optionally with a history-dependent variance head. The same-recipe comparisons of Section 6.2 vary only the factor family.

**Standing assumptions and identifiability target.** (A0) The observation-noise law is fixed and shared by every compared parameter set. The noise $\varepsilon_v^{(t)}$ is independent across sites and times and independent of $Z$, with a characteristic function that is nonzero almost everywhere. Every identifiability result below takes (A0), a known noise law shared by all compared models, as a hypothesis. (A1) Each $g_v$ is a surjective $C^2$ diffeomorphism of $\mathbb{R}$. (A2) $\mathrm{rank}\, F = D$. (A3) The marginal law of $Z^{(t)}$ has support $\mathbb{R}^D$. Two parameter sets $\theta$ and $\hat{\theta} = (\hat{F}, \hat{g}, P_{\hat{Z}})$ are *observationally equivalent* when the joint law of $X^{(1:T)}$ agrees. The class is identifiable up to permutation, scaling and translation when this forces

$$Z^{(1:T)} \overset{d}{=} \Big(J_0 \hat{Z}^{(t)} + z_0\Big)_{t\in[T]} \tag{3}$$

for a monomial $J_0 = PS$ ($P$ a permutation matrix, $S$ an invertible diagonal matrix) and some $z_0 \in \mathbb{R}^D$. The shift $z_0$ belongs in the class, because $\hat{g}_v(y) := g_v(y + F_v^\top z_0)$ absorbs it.

## 4 Incidence-structured spatial factors

**Why a heat kernel is not the right lift.** The direct analogue of a radial basis factor on an incidence structure is a heat kernel. Take the normalized hypergraph Laplacian $\Delta = I - \mathrm{diag}(d)^{-1/2} HW\,\mathrm{diag}(\delta)^{-1} H^\top \mathrm{diag}(d)^{-1/2}$ of Zhou et al. (2006), where $\delta_e$ is the size of unit $e$, and set each factor column to a normalized column of $e^{-\gamma\Delta}$. Proposition 1 in Appendix A shows that this is a heat kernel family of the weighted clique expansion $A_{\mathrm{ce}}$, in line with the reduction of this Laplacian to a pairwise expansion (Agarwal et al., 2006; Chitra & Raphael, 2019). Two hypergraphs with the same clique expansion and degrees thus give identical factor families.

**Factors built from whole units.** ISF avoids this collapse by building each footprint from columns of the incidence matrix $H$, rather than from an operator on $A_{\mathrm{ce}}$ (Figure 2). A column of $H$ marks the sites of one unit, so a footprint can follow a membership as a whole.

**Definition 1** (Incidence-structured spatial factor)**.** Let $A = [a_1, \ldots, a_D] \in \mathbb{R}_{\geq 0}^{M\times D}$ be nonnegative unit-to-mode weights. The factor is

$$F = \bar{H}A = \mathrm{diag}(d)^{-1} HW A. \tag{4}$$

Column $d$ of $F$ is the degree-normalized weighted union of the units selected by $a_d$, so a site in several units of several modes has several latent parents. We call the units from which the columns of $F$ are built *atoms*, and we restrict each atom to at most one mode.

**Definition 2** (Unit-level single-parent class)**.** The class $\mathcal{M}_{\mathrm{atom}}$ consists of factors of the form (4) with $\operatorname{rank} F = D$ for which the columns of $A$ have pairwise disjoint nonempty supports. Equivalently there is an assignment $\sigma : E \to \{0, 1, \ldots, D\}$ and amplitudes $\alpha_e > 0$ with $A_{ed} = \alpha_e \mathbf{1}[\sigma(e) = d]$, every mode receives at least one unit, and (A2) holds.

The restriction acts on units, and sites remain multi-parent, as the identifiability theory requires. With hyperedge atoms, Definition 1 depends on $H$ and not only on $A_{\mathrm{ce}}$, so it separates hypergraphs that the heat kernel cannot.

**Atoms and optional smoothing.** The atoms of ISF are the units of the incidence it receives: the hyperedges of a hypergraph, or the edges of a graph. Given only the pairwise operator, ISF receives the weighted clique expansion, whose atoms are the site pairs that co-occur in some unit. Section 6.3 uses it as the pair-level counterpart. An atom footprint has a hard boundary, so ISF can smooth each mode with one learned scale $\gamma_d > 0$,

$$F_{\cdot d} \;\propto\; e^{-\gamma_d \Delta}\, \bar{H} a_d. \tag{5}$$

The smoothing degenerates to the unsmoothed ISF as $\gamma_d \to 0$, and it spreads a sparse atom support into a connected region.

**Inference.** We implement a variational model in the style of Wang et al. (2025) (Appendix D.3) and compare factor families inside one shared encoder and graph learner. Unit assignments have a uniform categorical prior over $\{0, \ldots, D\}$. Training uses soft Gumbel-Softmax samples (Jang et al., 2017; Maddison et al., 2017) with a categorical KL term, and evaluation uses the hard argmax assignment. Amplitudes $\alpha_e = \mathrm{softplus}(\tilde{\alpha}_e)$ are deterministic learned parameters, and smoothing scales have Gaussian variational posteriors in log space. Without smoothing, factor multiplication costs $O(\mathrm{nnz}(H) \cdot D)$ per step. The heat kernel uses one dense eigendecomposition of $\Delta$, computed once in $O(L^3)$ time and $O(L^2)$ memory. Each application then costs $O(L^2 D)$. Appendix D.2 gives the tempered objective.

## 5 IDENTIFIABILITY

All proofs are in Appendices B and C. The results concern population identification of the latents for the unsmoothed atom class.

**Reduction to affine equivalence.** With identity observation maps, a spark condition on the dictionary of candidate columns suffices (Theorem 4 in Appendix B). This condition fails for the full atom family, whose members are linearly dependent. For the nonlinear case, two lemmas produce a $C^2$ diffeomorphism $\Theta$ with $Z^{(1:T)} \overset{d}{=} (\Theta(\hat{Z}^{(t)}))_t$ and

$$F_v^\top \Theta(\hat{z}) = h_v\left(\hat{F}_v^\top \hat{z}\right), \qquad h_v := g_v^{-1} \circ \hat{g}_v, \tag{6}$$

for every site $v$ and every $\hat{z}$. Call $F$ *decomposable* when the index set of its nonzero rows splits into nonempty $V_1$ and $V_2$ whose row spans meet only at the origin. Under (A0) to (A3) with $D \geq 2$ and $F$ indecomposable, Theorem 3 forces $\Theta(\hat{z}) = J_0 \hat{z} + z_0$ with $J_0$ invertible, makes every $h_v$ with $F_v \neq 0$ affine with slope $c_v \neq 0$, and gives $\hat{F} = C^{-1} F J_0$ with $C = \mathrm{diag}(c_v)$. An affine reduction of this kind was shown by Lyu & Fu (2022) for generic mixing matrices. We state it for indecomposable $F$ and prove it through Jacobian eigenvectors.

**Anchor and bridge conditions.** Theorem 3 leaves a linear algebra question. When $C\hat{F} = F J_0$ and both factors lie in one model class, must $J_0$ be monomial? For atom factors, two conditions on the incidence structure and its assignment suffice.

**Theorem 1** (Atom factors with site-specific observation maps)**.** *Let $\theta, \hat{\theta}$ with factors in $\mathcal{M}_{\mathrm{atom}}$ satisfy (A0) to (A3) with $D \geq 2$, and let the true $\theta$ satisfy:*

*(C1)* Anchor. *For every mode $d$ there is a unit $e$ with $\sigma(e) = d$ and a site $v \in e$ that belongs to no other unit.*

*(C2)* Bridge. *The mode co-occurrence graph is connected, where modes $d$ and $d'$ are adjacent when some site belongs both to a unit assigned to $d$ and to a unit assigned to $d'$.*

*Then observational equivalence forces* (3) *with a monomial $J_0$.*

The single-parent result of Brouillard et al. (2026) reaches a permutation and a coordinate-wise map in its own class. Theorem 1 reaches the narrower affine class under different assumptions. Section 6.3 examines anchored committee structures on which (C1) and (C2) are verified. Call two units adjacent when they share a site. Both conditions hold for every assignment in $\mathcal{M}_{\mathrm{atom}}$ that uses all units whenever each unit contains a site in no other unit and the resulting unit-overlap graph is connected (Appendix B). They name the features that pin down a latent representation: sites in one unit and sites shared by units of different modes.

**Transfer to the causal graph.** Proposition 2 (Appendix B) transfers Theorem 1 to the graph whenever the latent model lies in a class whose process law identifies its graph and which is closed under coordinate-wise affine maps.

### 5.1 Identification without anchors

Two further results need no anchor. The first, Theorem 2, concerns what a known incidence adds beyond its complete weighted pairwise operator $A_{\mathrm{ce}} = HW\,\mathrm{diag}(\delta)^{-1}H^{\top}$ (Proposition 1), diagonal included, under the population assumptions (A0) to (A3) with dependent latents allowed. A *legal* assignment is one of $\mathcal{M}_{\mathrm{atom}}$. Consider the unit-weight triples $E_A = \{012, 034, 135, 245\}$ and $E_B = \{013, 024, 125, 345\}$ on six labeled sites. Every site belongs to two triples. Set $D = M = 4$. Since every mode receives an atom, the class forces an atom-to-mode bijection, $F = H\,\mathrm{diag}(\alpha)P/2$ with $\alpha > 0$ and a permutation $P$, so every site has two latent parents and no row is pure.

**Theorem 2** (Incidence determines coordinates that pairwise input does not)**.** *In the class just specified, each fixed incidence identifies latent coordinates up to* (3) *under the population assumptions. Nevertheless, there exist models with these two incidences that share the full $A_{\mathrm{ce}}$, weighted degrees, unit sizes, and observed law, but whose latent laws are not related by any monomial affine transformation.*

Incidences with a common pairwise projection are known (Kim & Kirkland, 2022; Wang & Kleinberg, 2024). Theorem 2 shows that the difference reaches the latent coordinates. An explicit reflection certificate is in Appendix C.1. Theorem 2 concerns ISF with hyperedge atoms.

The second result, Theorem 5 in Appendix C.2, gives for every $m \geq 7$ an incidence with $2m$ sites and $m$ units in which every site belongs to three units, so no site is an anchor. Through the rank test of Li et al. (2017), the theorem proves that this incidence identifies the latents at $D = m - 1$ against every legal assignment.

## 6 Experiments

The experiments first evaluate the factor on a public grid benchmark, then compare unit-level with pair-level atoms and examine structures on which Theorem 1 applies.

### 6.1 Protocol

**Metrics.** Let $\rho_{dk}$ be the Pearson correlation between the true latent $Z_d$ and the estimated latent $\hat{Z}_k$ over all sequences and time points, and let $\hat{\pi} = \arg\max_{\pi} \sum_d |\rho_{d\pi(d)}|$ over permutations of $[D]$. The latent recovery score is $\mathrm{MCC} = D^{-1} \sum_d |\rho_{d\hat{\pi}(d)}|$. Let $\bar{G} \in [0,1]^{(\tau+1)\times D\times D}$ hold the posterior edge probabilities after relabeling by $\hat{\pi}$, and let $\Omega = \{(0,i,j) : i \neq j\} \cup \{(\ell,i,j) : \ell \in [\tau]\}$ be the candidate edges with true edge set $E^{\star} \subseteq \Omega$. With $\hat{E} = \{\omega \in \Omega : \bar{G}_{\omega} > 0.5\}$, the edge F1 is $2|\hat{E} \cap E^{\star}|/(|\hat{E}| + |E^{\star}|)$, and AUPRC is the average precision $|E^{\star}|^{-1} \sum_{k=1}^{|\Omega|} \mathrm{prec@}k \cdot \mathbf{1}[\omega_{(k)} \in E^{\star}]$ of the candidates $\omega_{(1)}, \omega_{(2)}, \ldots$ sorted by decreasing $\bar{G}_{\omega}$.

Table 1: Grid benchmark, twenty fresh draws per setting: mean edge F1 and AUPRC over draws, and draws won by our model against each method (F1, AUPRC; our row is the reference of *won*). Rows marked † report the native edge F1 of SPACY's code after its own latent matching, and their *won* compares our values under our evaluator with theirs. Their graph step is a lagged-dependence test that returns no instantaneous edges, and their AUPRC ranks the lagged candidates by its statistic (Appendix F.7). The released joint model was run on the first ten draws (Appendix F.1), on which our means are 0.669 / 0.681 ($D = 10$) and 0.651 / 0.653 ($D = 20$). *Calibrated*: its sparsity coefficient tuned with ground truth on a separate draw (Appendix F.6). Released code runs once per draw; *same recipe*: the cited factor family inside our encoder, graph learner and recipe.

| | $D = 10$ | | | $D = 20$ | | |
|---|---|---|---|---|---|---|
| Method | F1 | AUPRC | won | F1 | AUPRC | won |
| Mapped-PCMCI† (Tibau et al., 2022) | 0.361 | 0.390 | 20/20, 19/20 | 0.290 | 0.260 | 20/20, 20/20 |
| TPCA-PCMCI† (Babii et al., 2025) | 0.084 | 0.106 | 20/20, 20/20 | 0.042 | 0.050 | 20/20, 20/20 |
| LEAP† (Yao et al., 2022b) | 0.077 | 0.107 | 20/20, 20/20 | 0.044 | 0.051 | 20/20, 20/20 |
| TDRL† (Yao et al., 2022a) | 0.095 | 0.118 | 20/20, 20/20 | 0.048 | 0.053 | 20/20, 20/20 |
| SPACY, released (Wang et al., 2025) | 0.359 | 0.356 | 10/10, 10/10 | 0.326 | 0.279 | 10/10, 10/10 |
| calibrated | 0.319 | 0.316 | 9/10, 10/10 | 0.306 | 0.291 | 10/10, 10/10 |
| PCA-VARIMAX + our graph learner (Kaiser, 1958) | 0.608 | 0.637 | 14/20, 16/20 | 0.486 | 0.469 | 20/20, 20/20 |
| SINGLEPARENT, same recipe (Brouillard et al., 2026) | 0.492 | 0.489 | 20/20, 20/20 | 0.370 | 0.301 | 20/20, 20/20 |
| HEATKERNEL, same recipe (Laplacian of Zhou et al., 2006) | 0.597 | 0.617 | 17/20, 19/20 | 0.486 | 0.421 | 20/20, 20/20 |
| RBF, same recipe (Manning et al., 2014; Wang et al., 2025) | 0.631 | 0.619 | 14/20, 17/20 | 0.574 | 0.478 | 19/20, 20/20 |
| ISF (ours) | **0.666** | **0.699** | – | **0.657** | **0.662** | – |

**Unit of comparison.** A draw $r$ is one run of a generator with a fresh seed, which fixes the structure and the data. For a method $\mathcal{A}$ and a metric $\phi$, $\bar{\phi}_r(\mathcal{A}) = \frac{1}{3}\sum_{s=0}^{2} \phi_{r,s}(\mathcal{A})$ averages three model seeds fixed in advance (Appendix D). Our model wins draw $r$ against $\mathcal{B}$ when $\bar{\phi}_r(\text{ours}) > \bar{\phi}_r(\mathcal{B})$. With $K$ wins over $n$ draws, the exact two-sided sign test gives $p = \min\{1, 2\min(\Pr[Y \le K], \Pr[Y \ge K])\}$ with $Y \sim \text{Binomial}(n, 1/2)$. A tie counts as not won and stays in $n$. Whenever $K > n/2$, the resulting $p$ is at least the $p$ of the sign test that discards ties.

### 6.2 A PUBLIC GRID BENCHMARK

**Benchmark and methods.** We use the public synthetic generator of Wang et al. (2025) on a $100 \times 100$ grid, with $N = T = 100$, maximum lag $\tau = 2$, nonlinear latent dynamics with history-dependent spline noise and a linear observation map, at $D \in \{10, 20\}$, and generate twenty fresh draws per setting. Its ground-truth factors are radial basis kernels over site coordinates, the family of the coordinate kernel. ISF receives the grid adjacency as its incidence, so its atoms are the grid edges, and it uses the identity observation map. It smooths with (5) at $D = 20$ and not at $D = 10$, a configuration fixed before any fresh draw was generated. The coordinate kernel (RBF), the heat kernel (HEATKERNEL) and single-parent decoding (SINGLEPARENT) share our encoder, graph learner, recipe and calibrated sparsity $\lambda^\star$ (40 at $D = 10$, 150 at $D = 20$). PCA-VARIMAX fits our graph learner to fixed Varimax-rotated principal components. The released implementation of Wang et al. (2025) keeps its default configuration apart from training length (Appendix D.7) and is scored with our evaluator at its most favorable latent permutation. The two-stage and sequence-model baselines released with it are scored by their native edge F1 and by an AUPRC from the statistics of their graph step (Appendix F.7).

**Training recipe.** Four components of the training schedule do not refer to the factor family. Every factor family we train for Tables 1 and 2 receives the same components (Appendices D and E). The experiments of Section 6.3 use none of them (Appendix F.2). Let $\mathcal{L}_\lambda$ denote the training objective whose sparsity term weights the edge probabilities by $\lambda > 0$, and let $\hat{G}_\lambda(Z)$ be the graph that the graph learner fits to a latent sequence $Z$ under $\mathcal{L}_\lambda$.

*Sparsity calibration.* On a calibration draw with true latents $Z^\star$ and true graph $G^\star$, the oracle score $s(\lambda) = \text{F1}\big(\hat{G}_\lambda(Z^\star), G^\star\big)$ is evaluated on a finite grid $\Lambda_{\text{cal}}$. One value $\lambda^\star$ per setting is fixed as the grid value nearest to $0.4D^2$, a rule set after inspecting these scores (Appendix E), and every factor family we train receives $\lambda^\star$. The selection of $\lambda$ at the end of this subsection uses no ground truth.

Table 2: Ablations of ISF at $D = 20$ on the draws of Table 1 (same recipe, $\lambda = 150$): mean edge F1 and AUPRC, and draws won by the full ISF. The free variants also lack the assignment sampler and its KL term.

| Variant | F1 | AUPRC | Won by ISF |
|---|---|---|---|
| ISF without smoothing | 0.551 | 0.534 | 19/20, 20/20 |
| Free weights on grid edges, no assignment or smoothing | 0.572 | 0.586 | 19/20, 20/20 |
| Free footprint, no incidence, assignment or smoothing | 0.562 | 0.564 | 20/20, 20/20 |
| ISF (full) | **0.657** | **0.662** | – |

*Linear warm start.* During the first half of the graph schedule, every $f_d$ in (2) is linear.

*Graph relearning.* After the representation stage, encoder, factor and observation maps are frozen, and $(G, f)$ is relearned from scratch on the standardized posterior means $\bar{Z}$.

*Low-prior representation stage.* The representation is learned with prior weight $\beta = 0.05$ on the latent prior and entropy terms of the objective, and the graph is relearned at the default weight $\beta = D/4$.

**Results.** Our model attains the highest mean F1 and AUPRC in both settings and wins most paired draws against every baseline (Table 1; Appendix F). Against the coordinate kernel trained with the same recipe, whose family matches the generator's true factors, it improves mean F1 and AUPRC in both settings. The improvement is significant under the exact sign test in AUPRC in both settings and in F1 at $D = 20$ (Appendix F.11). The released joint model trails the same-recipe kernel, and the kernel trails our factor. The first gap reflects the recipe and implementation. The second, under a fixed recipe, reflects the factor family. The same-recipe heat kernel and single-parent decoding also trail our factor (Appendix F.13). The released two-stage and sequence-model baselines fall below our model in both settings (Appendix F.7).

**Ablations.** Table 2 removes parts of ISF at $D = 20$, where all of them are active (Appendices F.12 and G). The full ISF outperforms each variant on almost every draw in F1 and on every draw in AUPRC. When a degree- and size-preserving rewiring replaces the grid adjacency, with the number of atoms, the recipe and $\lambda$ kept, our factor falls below the coordinate kernel on every tested draw in both settings (Appendix F.8). The gain therefore requires the correct incidence. Appendix G lists further ablations, including those on committee hypergraphs. In the parameter study of Appendix H, ISF has a higher mean F1 and AUPRC than the same-recipe coordinate kernel at every tested $\lambda$ in both settings.

**Sparsity selected without ground truth.** Let $\bar{z}_{n,t}$ be the standardized encoder means of the held-out sequences $n \in \mathcal{N}_{\mathrm{v}}$ (the released implementation's validation sequences), and let the model be trained on the remaining sequences for every $\lambda$ in the grid $\Lambda = \{5, 10, 20, 40, 80, 160, 320\}$, fixed in advance. We select

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \; -\frac{1}{|\mathcal{I}_{\mathrm{v}}|} \sum_{(n,t) \in \mathcal{I}_{\mathrm{v}}} \sum_{d=1}^{D} \log p_{\hat{G}_\lambda, \hat{f}_\lambda} \left( \bar{z}_{n,t,d} \mid \mathrm{Pa}^d_{\hat{G}_\lambda}(< t),\, \mathrm{Pa}^d_{\hat{G}_\lambda}(t) \right), \tag{7}$$

where $\mathcal{I}_{\mathrm{v}} = \{(n,t) : n \in \mathcal{N}_{\mathrm{v}},\ \tau < t \le T\}$ and $p_{\hat{G}_\lambda, \hat{f}_\lambda}$ is the structural model (2) with the fitted functions and the fitted edge probabilities as parent weights (Appendix F.9). The score contains no sparsity term and no ground truth. With $\hat{\lambda}$, our model outperforms the released joint model on every draw, whether that model uses its default or its calibrated sparsity. When both factors use this selection, our factor retains a higher mean F1 and AUPRC than the coordinate kernel in both settings (Appendix F.9). When the coordinate kernel instead selects its own representation settings on separate calibration draws, our factor still outperforms it (Appendix F.10).

### 6.3 Unit-level and pair-level atoms on committee hypergraphs

We examine whether reading memberships as whole units improves latent recovery over pair atoms and the heat kernel built on the same weighted clique expansion. We use random committee hyper-

Table 3: Random committee hypergraphs, ten fresh structures at two noise levels: mean latent MCC, edge F1 and AUPRC. Shaded rows are ours: ISF on the true memberships, and ISF on memberships reconstructed from $A_{\mathrm{ce}}$ in a separate experiment (Appendix F.3). The other ISF rows are controls, and the table uses no bold. ISF on the true memberships has a higher MCC than each of the three comparators on every structure at both noise levels (Table 9).

| | noise 2.0 | | | noise 1.5 | | |
|---|---|---|---|---|---|---|
| Method | MCC | F1 | AUPRC | MCC | F1 | AUPRC |
| ISF on rewired memberships | 0.348 | 0.083 | 0.150 | 0.419 | 0.105 | 0.152 |
| HEATKERNEL (same pairwise operator) | 0.464 | 0.099 | 0.161 | 0.570 | 0.143 | 0.181 |
| ISF on the clique expansion (pair atoms) | 0.702 | 0.411 | 0.388 | 0.839 | 0.640 | 0.629 |
| ISF on memberships reconstructed from $A_{\mathrm{ce}}$ | 0.801 | 0.536 | 0.497 | 0.882 | 0.698 | 0.669 |
| ISF on the true memberships | 0.791 | 0.537 | 0.492 | 0.883 | 0.689 | 0.665 |

graphs ($L = 300$ sites, $M = 60$ hyperedges of 20 to 40 members, overlapping memberships) whose spatial footprints follow the memberships. Each true footprint is a union of hyperedges, as in ISF. We fit each of ten fresh structures at observation-noise standard deviations $2.0$ and $1.5$ ($N = 100$, three model seeds). We compare ISF on the true memberships with ISF on their clique expansion (pair atoms), the heat kernel on the same pairwise operator (HEATKERNEL), and ISF on memberships rewired by degree- and size-preserving swaps. We also fit ISF on memberships reconstructed from the weighted clique expansion alone.

ISF on the true memberships achieves a higher MCC than each of the three comparators on every structure at both noise levels. Its mean edge F1 and AUPRC are higher as well (Table 3; Appendix F.2). The two ISF controls in Table 3 ablate the input of ISF: they reduce the units to pairs or misplace them. The rewiring control keeps the degree and size sequences and the number of atoms. Its deficit therefore shows that the gain depends on which sites share a unit, not on the number or the sizes of the atoms. A symmetric nonnegative factorization of the weighted clique expansion recovers every hyperedge of these structures exactly. ISF on the reconstructed memberships outperforms ISF on the clique expansion on almost every structure and performs as on the true memberships (Appendix F.3). The advantage therefore lies in the hyperedge-atom dictionary, which on these structures is available from the pairwise operator alone.

**Structures on which Theorem 1 applies.** We further generate ten committee structures with $L = 450$ sites in which every mode owns an anchor site, verify (C1) and (C2) on each, and keep most sites multi-parent (Appendix F.4). At observation-noise standard deviation $0.5$, the unsmoothed ISF estimator in the class of Theorem 1 recovers the latents almost exactly on almost every structure, consistent with the population statement of the theorem. We then observe every site $v$ through its own map $g_v(y) = y + \mu_v \tanh(\nu_v y + \zeta_v) - \mu_v \tanh(\zeta_v)$ with $|\mu_v \nu_v| < 0.9$, a $C^2$ diffeomorphism of $\mathbb{R}$ as in (A1). The estimator learns a map of the same form for every site, with its parameters constrained so that the map stays invertible (Appendix F.5). With these maps learned jointly, the estimator still recovers the latents accurately on every anchored structure, and more accurately than the same estimator with an identity map (Appendix F.5).

## 7 CONCLUSION

The pairwise operator of a hypergraph does not in general determine the latent coordinates that its incidence determines. Incidence-structured spatial factors build footprints from whole units, and for the unsmoothed family anchor and bridge sites identify the latents up to permutation, scaling and translation. Some incidences identify them even without anchor sites. On a public grid benchmark, the factor recovers latent causal graphs more accurately than kernel, single-parent, two-stage and sequence-model baselines, with a gain that requires the correct incidence. On random committee hypergraphs with membership-following footprints, unit-level atoms, given or reconstructed from the weighted clique expansion, outperform pair-level atoms and heat kernels. Latent representations of data whose footprints follow known memberships can therefore benefit from spatial factors built on whole units.

### Reproducibility statement

The model, the generative process and the standing assumptions are specified in Sections 3 and 4, the training recipe in Section 6.2 and the evaluation protocol in Section 6.1. Proofs of every stated theorem and proposition, with the assumptions they use, are in Appendices A to C. Appendix D documents the hyperparameters, the seeds and how they were chosen, the data generation and the execution of the released baseline code. Appendix E gives the sparsity calibration, Appendix F the records of the confirmation experiments, with summary statistics for the reported comparisons and per-draw values for the principal ones, and Appendix J the compute. The supplementary material contains the code of our model and of the same-recipe coordinate kernel, and the scripts that run their experiments.

### Ethics statement

All experiments use synthetic signals generated by public benchmark code or by our own generators. No human-subject or individual-level data are involved.

### AI use statement

As guided by the conference AI disclosure policy, in this work, we used generative AI tools to assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat datasets, formulate mathematical claims and support qualitative and thematic data analysis. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, interpret results, propose or refine hypotheses. Additionally, we also used generative AI tools to search related works, formulate questions for surveys or interviews, help authors create or modify scientific figures or images, help authors build experiment pipelines, help improve the writing quality and clarity of the human-drafted manuscript, edit a research paper to improve readability, sourcing/searching for information and check for potential plagiarism, brainstorming, identify relevant literature, format references. All work is first drafted by human authors and subsequently reviewed and edited by human authors, with the help of generative AI tools. And all AI-assisted work has been reviewed and verified by human authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## References

Sophie Achard and Christian Jutten. Identifiability of post-nonlinear mixtures. *IEEE Signal Processing Letters*, 12(5):423–426, 2005. doi: 10.1109/LSP.2005.845593.

Sameer Agarwal, Kristin Branson, and Serge Belongie. Higher order learning with graphs. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 17–24. ACM, 2006.

Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization – provably. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, pp. 145–162. ACM, 2012.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 280–288. PMLR, 2013.

Andrii Babii, Eric Ghysels, and Junsu Pan. Tensor PCA for factor models. *Journal of Econometrics*, pp. 106077, 2025. doi: 10.1016/j.jeconom.2025.106077.

Xin Bing, Florentina Bunea, Yang Ning, and Marten H. Wegkamp. Adaptive estimation in structured factor models with applications to overlapping clustering. *The Annals of Statistics*, 48(4):2055–2081, 2020.

Philippe Brouillard, Sébastien Lachapelle, Julia Kaltenborn, Yaniv Gurwicz, Dhanya Sridhar, Alexandre Drouin, Peer Nowack, Jakob Runge, and David Rolnick. Learning a spatial partitioning and its causal relations from temporal data. In *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pp. 547–592. PMLR, 2026.

Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Uthsav Chitra and Benjamin J. Raphael. Random walks on hypergraphs with edge-dependent vertex weights. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1172–1181, 2019.

Chris Ding, Xiaofeng He, and Horst D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pp. 606–610, 2005.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, volume 16, 2003.

David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell_1$ minimization. *Proceedings of the National Academy of Sciences*, 100(5): 2197–2202, 2003. doi: 10.1073/pnas.0437847100.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3558–3565, 2019.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *International Conference on Learning Representations*, 2023.

Matthias Hein, Simon Setzer, Leonardo Jost, and Syama Sundar Rangapuram. The total variation on hypergraphs - learning on hypergraphs revisited. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017.

Henry F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23 (3):187–200, 1958. doi: 10.1007/BF02289233.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217, 2020.

Sooyeong Kim and Steve Kirkland. Gram mates, sign changes in singular values, and isomorphism. *Linear Algebra and its Applications*, 644:108–148, 2022.

Yanjun Li, Kiryung Lee, and Yoram Bresler. Identifiability in bilinear inverse problems with applications to subspace or sparsity-constrained blind gain and phase calibration. *IEEE Transactions on Information Theory*, 63(2):822–842, 2017. doi: 10.1109/TIT.2016.2637933. Preprint: arXiv:1501.06120.

Qi Lyu and Xiao Fu. Provable subspace identification under post-nonlinear mixtures. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1554–1567. Curran Associates, Inc., 2022.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Jeremy R. Manning, Rajesh Ranganath, Kenneth A. Norman, and David M. Blei. Topographic factor analysis: A Bayesian model for inferring brain networks from neural data. *PLoS ONE*, 9 (5):e94914, 2014.

Weiguang Mao, Elena Zaslavsky, Boris M. Hartmann, Stuart C. Sealfon, and Maria Chikina. Pathway-level information extractor (PLIER) for gene expression data. *Nature Methods*, 16(7): 607–610, 2019.

Gemma E. Moran, Dhanya Sridhar, Yixin Wang, and David M. Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022. URL `https://openreview.net/forum?id=vd0onGWZbE`.

Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36249–36293. PMLR, 2024.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1388–1397. PMLR, 2020.

Jakob Runge, Vladimir Petoukhov, Jonathan F. Donges, Jaroslav Hlinka, Nikola Jajcay, Martin Vejmelka, David Hartman, Norbert Marwan, Milan Paluš, and Jürgen Kurths. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, 6:8502, 2015. doi: 10.1038/ncomms9502.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11):eaau4996, 2019. doi: 10.1126/sciadv.aau4996.

Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

Nils Sturma, Miriam Kranzlmueller, Irem Portakal, and Mathias Drton. Matching criterion for identifiability in sparse factor analysis. *Psychometrika*, 91(2):536–555, 2026.

Xavier-Andoni Tibau, Christian Reimers, Andreas Gerhardus, Joachim Denzler, Veronika Eyring, and Jakob Runge. A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections. *Environmental Data Science*, 1:e12, 2022. doi: 10.1017/eds.2022.11.

Kun Wang, Sumanth Varambally, Duncan Watson-Parris, Yian Ma, and Rose Yu. Discovering latent causal graphs from spatiotemporal data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64549–64585. PMLR, 2025.

Yanbang Wang and Jon M. Kleinberg. From graphs to hypergraphs: Hypergraph projection and its reconstruction. In *International Conference on Learning Representations (ICLR)*, 2024.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14891–14902, 2020.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26492–26503. Curran Associates, Inc., 2022a.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022b.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16411–16422. Curran Associates, Inc., 2022.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

## A CLIQUE-EXPANSION COLLAPSE

**Proposition 1** (Clique-expansion collapse)**.** *Write* $A_{\mathrm{ce}} = HW\mathrm{diag}(\delta)^{-1}H^{\top}$. *Then* $\Delta = I - \mathrm{diag}(d)^{-1/2}A_{\mathrm{ce}}\mathrm{diag}(d)^{-1/2}$*, so the heat kernel family of* $\Delta$ *is a heat kernel family of the weighted clique expansion. Two hypergraphs with the same* $A_{\mathrm{ce}}$ *and the same degrees give identical factor families, and two such hypergraphs with different incidence matrices exist.*

The identity for $\Delta$ is immediate from the definition of the normalized hypergraph Laplacian. For the second part, let $V = \{1, 2, 3\}$. Let $\mathcal{H}_1$ carry the unit $\{1, 2, 3\}$ with weight $w$ together with three singleton units of weight $w/3$ each. Let $\mathcal{H}_2$ carry the three pairs with weight $2w/3$ each. In both cases every off-diagonal entry of $A_{\mathrm{ce}}$ equals $w/3$, every diagonal entry equals $2w/3$, and every degree equals $4w/3$. The normalized Laplacians therefore agree while the incidence matrices differ.

The collapse argument therefore motivates building the factor from the incidence matrix itself.

## B PROOFS

This appendix holds the proofs of the results in Section 5 other than those of Section 5.1, together with the two preliminary lemmas. For the linear case, a dictionary of candidate factor columns is a family $\mathcal{K} = \{k_\psi \in \mathbb{R}^L\}$ indexed injectively by $\psi$, and its *spark* is the smallest number of distinct members that are linearly dependent (Donoho & Elad, 2003). Theorem 4 needs any $D + 1$ members to be independent, where the linear-case theorem of Wang et al. (2025) assumes a linearly independent family and uses the independence of up to $2D$ members.

**Theorem 3** (Affine reduction; cf. Lyu & Fu, 2022)**.** *Assume (A0) to (A3), let* $D \geq 2$*, let* $F$ *be indecomposable, and let* $\theta$ *and* $\hat{\theta}$ *be observationally equivalent. Then* $\Theta(\hat{z}) = J_0\hat{z} + z_0$ *for an invertible* $J_0$*, each* $h_v$ *with* $F_v \neq 0$ *is affine with slope* $c_v \neq 0$*, and* $\hat{F} = C^{-1}FJ_0$ *with* $C = \mathrm{diag}(c_v)$.

**Theorem 4** (Linear case)**.** *Assume (A0), (A2) and (A3), let* $g_v = \hat{g}_v = \mathrm{id}$*, and let the columns of* $F$ *and* $\hat{F}$ *come from one dictionary* $\mathcal{K}$ *with* $\mathrm{spark}(\mathcal{K}) > D + 1$*. Observational equivalence then forces* $\hat{F} = FP$ *for a permutation matrix* $P$*, and* $\hat{Z}^{(1:T)} \overset{d}{=} P^{\top}Z^{(1:T)}$.

**Lemma 1** (Denoising)**.** *Under (A0), observational equivalence of* $\theta$ *and* $\hat{\theta}$ *implies that the laws of the noiseless mixtures agree,* $(m_\theta(Z^{(t)}))_t \overset{d}{=} (m_{\hat{\theta}}(\hat{Z}^{(t)}))_t$.

*Proof.* Write $Y = (m_\theta(Z^{(t)}))_t$ and define $\hat{Y}$ analogously, both valued in $\mathbb{R}^{LT}$. As in the proof of Khemakhem et al. (2020, Theorem 1), independence of the noise gives $\varphi_Y\varphi_\varepsilon = \varphi_{\hat{Y}}\varphi_\varepsilon$ for the characteristic functions. The factor $\varphi_\varepsilon$ is a product of coordinate characteristic functions and is nonzero almost everywhere, so $\varphi_Y = \varphi_{\hat{Y}}$ almost everywhere, hence everywhere by continuity. □

**Lemma 2** (Reparameterization map)**.** *Let* $\theta$ *and* $\hat{\theta}$ *satisfy (A0) to (A3) and be observationally equivalent. Then there is a* $C^2$ *diffeomorphism* $\Theta$ *of* $\mathbb{R}^D$ *with* $Z^{(1:T)} \overset{d}{=} (\Theta(\hat{Z}^{(t)}))_t$*, and* (6) *holds for every site* $v$ *and every* $\hat{z}$.

*Proof.* Assumptions (A1) and (A2) make $m_\theta$ injective. If $\|z\| \to \infty$ then $\|Fz\| \to \infty$, so some $|g_v(F_v^{\top}z)| \to \infty$, and $m_\theta$ is proper with closed image. The support of a pushforward under a continuous map is the closure of the image of the support, so (A3) gives $\mathrm{supp}\,\mathcal{L}(m_\theta(Z^{(t)})) = m_\theta(\mathbb{R}^D)$. Lemma 1 equates the two supports, so $m_\theta(\mathbb{R}^D) = m_{\hat{\theta}}(\mathbb{R}^D)$ and $\Theta := m_\theta^{-1} \circ m_{\hat{\theta}}$ is a

bijection. Choose $S \subset V$ with $|S| = D$ and $F_S$ invertible. Then $\Theta(\hat{z}) = F_S^{-1}(h_v(\hat{F}_v^\top \hat{z}))_{v \in S}$, which is $C^2$, and symmetrically for $\Theta^{-1}$. Applying $g_v^{-1}$ componentwise to $m_\theta \circ \Theta = m_{\hat{\theta}}$ gives (6). The equality of laws follows from injectivity of $m_\theta$ and Lemma 1. □

**Proof of Theorem 4.** Under identity observation maps, (6) reads $F\Theta(\hat{z}) = \hat{F}\hat{z}$, so $\Theta$ is linear. Write it as $R$, so that $\hat{F} = FR$ with $R$ invertible. Column $i$ gives $k_{\hat{\psi}_i} = \sum_j R_{ji} k_{\psi_j}$. If $\hat{\psi}_i$ were not among the $\psi_j$ then $D+1$ distinct dictionary members would be dependent, contradicting the spark condition. Hence $\hat{\psi}_i = \psi_{\pi(i)}$, and independence of $k_{\psi_1}, \ldots, k_{\psi_D}$ forces $R_{\cdot i} = e_{\pi(i)}$. Distinctness of the $\hat{\psi}_i$ makes $\pi$ a permutation. □

**Proof of Theorem 3.** Let $J(\hat{z})$ be the Jacobian of $\Theta$ and put $c_v(\hat{z}) = h_v'(\hat{F}_v^\top \hat{z}) \neq 0$. Differentiating (6) gives

$$F_v^\top J(\hat{z}) = c_v(\hat{z})\, \hat{F}_v^\top \quad (\forall v) \qquad \Longleftrightarrow \qquad \hat{F} = C(\hat{z})^{-1} F J(\hat{z}). \tag{8}$$

Fix $\hat{z}_1, \hat{z}_2$ and put $R = J(\hat{z}_1)J(\hat{z}_2)^{-1}$ and $\lambda_v = c_v(\hat{z}_1)/c_v(\hat{z}_2)$. Then $F_v^\top R = \lambda_v F_v^\top$, so every nonzero row of $F$ is a real eigenvector of $R^\top$. Those rows span $\mathbb{R}^D$, so $R^\top$ is diagonalizable with $\mathbb{R}^D = \mathcal{V}_1 \oplus \cdots \oplus \mathcal{V}_k$ and every nonzero row in exactly one eigenspace. If $k \geq 2$ then grouping rows by eigenspace decomposes $F$, which is excluded. Hence $R = \lambda I$, so $J(\hat{z}) = \kappa(\hat{z})J(0)$ and $c_v(\hat{z}) = \kappa(\hat{z})c_v(0)$ for a scalar function $\kappa$.

For each nonzero row, $\kappa(\hat{z}) = h_v'(\hat{F}_v^\top \hat{z})/h_v'(0)$ depends on $\hat{z}$ only through $\hat{F}_v^\top \hat{z}$, and $\kappa \in C^1$, so $\nabla\kappa(\hat{z}) \in \operatorname{span}(\hat{F}_v)$. By (8), $F_v \neq 0$ implies $\hat{F}_v \neq 0$, and $\operatorname{rank} \hat{F} = D \geq 2$ supplies two nonparallel rows $\hat{F}_v$ and $\hat{F}_{v'}$. Their spans meet only at the origin, so $\nabla\kappa \equiv 0$, and connectedness of $\mathbb{R}^D$ gives $\kappa \equiv \kappa(0) = 1$. Therefore $J \equiv J_0 := J(0)$ and $\Theta$ is affine. Each $h_v'$ is constant on $\{\hat{F}_v^\top \hat{z}\} = \mathbb{R}$, so $h_v$ is affine, and (8) gives $\hat{F} = C^{-1} F J_0$. □

**Proof of Theorem 1.** First, $F$ is indecomposable. An anchor site $v \in e$ with $\sigma(e) = d$ has $F_v = \bar{H}_{ve}\alpha_e \mathbf{e}_d \propto \mathbf{e}_d$, so every basis vector $\mathbf{e}_d$ appears among the rows up to scale. Suppose $V_1, V_2$ decomposed $F$ with spans $U_1, U_2$. The rows span $\mathbb{R}^D$, so $U_1 \oplus U_2 = \mathbb{R}^D$. Each $\mathbf{e}_d$ lies in one summand, and counting dimensions gives $U_i = \operatorname{span}\{\mathbf{e}_d : d \in S_i\}$ for a nonempty partition $S_1 \sqcup S_2 = [D]$. Every nonzero row then has its support inside $S_1$ or inside $S_2$, so the mode co-occurrence graph is disconnected, contradicting (C2).

Theorem 3 now gives $C\hat{F} = FJ_0$. For rigidity, take a mode $d$ and an anchor site $v \in e$ with $\sigma(e) = d$. Since $v$ lies in no other unit, $\hat{F}_v^\top = \bar{H}_{ve}\hat{A}_{e,\cdot}$ is either zero or proportional to $\mathbf{e}_{\hat{\sigma}(e)}^\top$. The right-hand side is $F_v^\top J_0 = \bar{H}_{ve}\alpha_e (J_0)_{d,\cdot} \neq 0$ because $J_0$ is invertible, so the left-hand side is nonzero and $(J_0)_{d,\cdot} \propto \mathbf{e}_{\hat{\sigma}(e)}^\top$. This holds for every $d$, so every row of $J_0$ has exactly one nonzero entry. Invertibility makes $J_0$ monomial. □

**Proposition 2** (Conditional transfer to the causal graph)**.** *Suppose* (3) *holds with a monomial $J_0$, so that $\hat{Z}_i = a_i Z_{\pi(i)} + b_i$ with $a_i \neq 0$ in law, as granted by Theorem 1. Suppose further that the latent processes of both parameter sets follow temporal structural causal models from a class $\mathcal{S}$ such that (i) within $\mathcal{S}$ the law of the latent process determines the graph up to relabelling of nodes, and (ii) $\mathcal{S}$ is closed under coordinate-wise affine maps with nonzero slopes combined with a relabelling of nodes. Then $\hat{G}$ equals $G$ up to the node permutation $\pi$.*

*Proof.* By (ii), the process $(a_i Z_{\pi(i)} + b_i)_i$ follows a model in $\mathcal{S}$ whose graph is $G$ relabelled by $\pi$. It has the law of $\hat{Z}$, which by assumption follows a model in $\mathcal{S}$ with graph $\hat{G}$. By (i) the two graphs agree. □

For the class of Gong et al. (2023), we sketch why (ii) holds, without restating their theorem. Their Assumptions 1 to 3 concern conditional dependence, conditional independence and minimality, which a coordinate-wise affine map with nonzero slopes preserves in both directions. Assumption 4 concerns the variable set, and an invertible reparameterization introduces no new common cause. Assumption 5 requires an absolutely continuous joint law with finite log density, and the transformed log density shifts by a finite constant per time step. Their proof further uses third-order

differentiability, which composition with an affine map preserves. It also uses a condition that the two orientations of an instantaneous pair induce different conditional log densities, which an affine change of variables leaves unchanged.

**Sufficient condition of Section 5.** Two units are adjacent in the *unit-overlap graph* when they share a site. Let every unit contain a site that belongs to no other unit, let the unit-overlap graph be connected, and let $\sigma$ be a legal assignment with $\sigma(e) \neq 0$ for every unit $e$. For (C1), every mode $d$ receives a unit $e$ with $\sigma(e) = d$ (Definition 2), and $e$ contains a site that belongs to no other unit. For (C2), take modes $d \neq d'$, units $e, e'$ with $\sigma(e) = d$ and $\sigma(e') = d'$, and a path $e = e_0, e_1, \ldots, e_k = e'$ in the unit-overlap graph. Consecutive units $e_{i-1}, e_i$ share a site, which belongs to a unit assigned to $\sigma(e_{i-1})$ and to a unit assigned to $\sigma(e_i)$. These two modes are therefore equal or adjacent in the mode co-occurrence graph, so $d$ and $d'$ are connected in it. □

## C Proofs for memberships beyond the pairwise operator

This appendix proves Theorems 2 and 5. We write $A_{\mathrm{ce}}$ for the complete weighted pairwise operator of Proposition 1, including its diagonal. The conclusion of Theorem 3 is written $\mathrm{diag}(c)\widehat{F} = FJ$, with $\mathrm{diag}(c) = C$ and $J = J_0$. A matrix is monomial when it is a permutation times an invertible diagonal matrix, and a *legal* assignment is a positive unit-single-parent assignment of $\mathcal{M}_{\mathrm{atom}}$ with every mode supported. The population assumptions are (A0) to (A3).

### C.1 The six-site twin

With $u = (1, -1, -1, 1)^\top$, define

$$O = I - \tfrac{1}{2}uu^\top, \qquad O^2 = I, \qquad H_A O = H_B. \tag{9}$$

Thus $H_A H_A^\top = H_B H_B^\top$ and the two operators $A_{\mathrm{ce}}$ coincide. Unit amplitudes and $Z_B = OZ_A$ give $F_B Z_B = F_A Z_A$ for every $Z_A$, before applying the same site maps and adding the same noise. Taking $Z_A \sim \mathcal{N}(0, \mathrm{diag}(1, 2, 3, 5))$ gives a non-diagonal covariance for $Z_B$. A monomial transform of a diagonal-covariance Gaussian still has diagonal covariance, proving the claimed difference of latent laws. Independent-component priors would exclude this alternative; dependence is an explicit part of the comparison class.

For fixed $H$, the two overlapping minimal row dependencies $r_0 + r_5 = r_1 + r_4 = r_2 + r_3$ establish indecomposability. Any diagonal row rescaling preserving the column space must be constant on each circuit, hence constant on all six rows. Applying Theorem 3 and cancelling the full-rank dictionary leaves only a monomial $J$. The matrices and a direct proof follow. Exhaustive reconstruction from $A_{\mathrm{ce}}$ in the same unit-weight four-triple class returns exactly $E_A$ and $E_B$; a pairwise-input method may correctly return this ambiguity set.

The two unnormalized incidence matrices and reflection are

$$H_A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad H_B = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad O = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}.$$

Direct multiplication verifies (9). Each listed four-row circuit has rank three, and every three-row subset has rank three. Their union covers all rows and they overlap, proving indecomposability. A four-row minor is nonzero, proving rank four. In the same-$H$ bijection class, both loading column spaces equal $\mathrm{col}(H)$. If $\mathrm{diag}(c)\widehat{F} = FJ$, this diagonal map preserves $\mathrm{col}(H)$. Applying it to the unique circuit relations forces all $c_v$ equal. Cancelling $H$ yields $J = c\,P^{-1}\,\mathrm{diag}(\alpha)^{-1}\,\mathrm{diag}(\widehat{\alpha})\widehat{P}$ with the appropriate assignment permutations, hence a monomial matrix. This conclusion allows unknown positive amplitudes.

The alternate covariance is

$$O\,\mathrm{diag}(1,2,3,5)O^\top = \frac{1}{4}\begin{pmatrix} 11 & -5 & -3 & -1 \\ -5 & 11 & 1 & 3 \\ -3 & 1 & 11 & 5 \\ -1 & 3 & 5 & 11 \end{pmatrix}.$$

The off-diagonal entries establish a law-level distinction, not just a different parameterization of an isotropic Gaussian. For any coordinate permutation, the sum of squared cross-correlations between $Z_A$ and $Z_B$ is one. The average residual of any one-to-one coordinate matching is therefore $1 - 1/4 = 3/4$ in either direction. This is a mismatch property of these two representations, not a uniform risk bound on all estimators receiving $A_{\mathrm{ce}}$.

### C.2 The cyclic family

**Theorem 5** (A cyclic fixed-incidence family at $D = m - 1$)**.** *For every $m \geq 7$, index $m$ units by $\mathbb{Z}_m$ and define a distinct $2m \times m$ incidence $H_m^{\mathrm{cyc}}$ with site rows*

$$H_{C_i,:} = e_i^\top + e_{i+1}^\top + e_{i+2}^\top, \qquad H_{Q_i,:} = e_i^\top + e_{i+1}^\top + e_{i+3}^\top. \tag{10}$$

*Set $D = m-1$ and take* any *legal positive unit-single-parent true assignment $A \in \mathbb{R}_{\geq 0}^{m\times(m-1)}$, with every mode supported. Compare it against every legal assignment on the same labeled $H_m^{\mathrm{cyc}}$, allowing different positive unit weights. Under the population assumptions, observational equivalence implies* (3)*. This includes every positive-amplitude merged-pair assignment and every one-zero-unit assignment on both sides. Every loading in the comparison class has rank $m-1$, a connected row matroid (it is indecomposable in the sense of Section 5), no zero rows and at least two modes at each site.*

**A complete incidence annihilator.** Write $H = H_m^{\mathrm{cyc}}$ and use all indices modulo $m$. To distinguish sites from incidence rows, let $E_{C_i}, E_{Q_i} \in \mathbb{R}^{2m}$ be the standard basis vectors of the *site* space. Define $R \in \mathbb{R}^{m\times 2m}$ by

$$R_{i,:} = E_{C_i}^\top + E_{C_{i+1}}^\top + E_{C_{i+3}}^\top - E_{Q_i}^\top - E_{Q_{i+1}}^\top - E_{Q_{i+2}}^\top. \tag{11}$$

Both the positive and negative sums after multiplication by $H$ equal $e_i^\top + 2e_{i+1}^\top + 2e_{i+2}^\top + 2e_{i+3}^\top + e_{i+4}^\top + e_{i+5}^\top$, so $RH = 0$. These six offsets are consecutive residues modulo $m$, hence distinct for every $m \geq 7$. If $Hx = 0$, subtracting the $C_i$ equation from the $Q_i$ equation gives $x_{i+3} = x_{i+2}$ for every $i$. Thus $x$ is constant and a $C_i$ equation makes it zero: $\mathrm{rank}\, H = m$. If $u^\top R = 0$, its $C_j$ and $Q_j$ coefficients respectively give $u_j + u_{j-1} + u_{j-3} = 0$ and $u_j + u_{j-1} + u_{j-2} = 0$. Subtraction makes $u$ constant, and either equation makes it zero. Hence $\mathrm{rank}\, R = m$ and $\ker R = \mathrm{col}(H)$.

**Every legal assignment and the rank gate.** Let $A \in \mathbb{R}_{\geq 0}^{m\times(m-1)}$ be legal, with every mode nonempty. The number of active unit rows is $m-1$ or $m$. In the former case exactly one unit is zero and all active units support distinct modes; in the latter exactly one pair shares a mode and all other units are singletons. Let $S_A$ be the singleton units. Thus at most two units lie outside $S_A$. For each $j \in S_A$, a column of $HA$ equals $\alpha_j H_{\cdot j}$ with $\alpha_j > 0$.

Define the linear map $\mathcal{K}_A : c \mapsto R\,\mathrm{diag}(c)HA$ and put $a_i = c_{C_i}$, $b_i = c_{Q_i}$. If $\mathcal{K}_A(c) = 0$, then $R_{i,:}\,\mathrm{diag}(c)H_{\cdot j} = 0$ for every $j \in S_A$ and every $i$. For any such $j$, five annihilator rows give

| $i$ | equation |
|---|---|
| $j$ | $a_j = b_j$ |
| $j-5$ | $a_{j-2} = b_{j-3}$ |
| $j-4$ | $a_{j-1} = b_{j-3}$ |
| $j-3$ | $a_{j-2} + a_j = b_{j-3} + b_{j-1}$ |
| $j-2$ | $a_{j-2} + a_{j-1} = b_{j-1} + b_j$ |

(12)

The first three equations give $b_j = a_j$ and $a_{j-2} = a_{j-1} = b_{j-3}$. The fourth gives $b_{j-1} = a_j$, and the fifth gives $a_{j-2} = a_j$. These are exactly the six sites incident to $j$:

$$a_{j-2} = a_{j-1} = a_j = b_{j-3} = b_{j-1} = b_j.$$

For every $m \geq 7$ the relation rows $j-5, \ldots, j$ are distinct residues modulo $m$, as are the six unit offsets, so there is no wraparound exception. To propagate the local value $\gamma_j$, connect two units when they share a $C_i$ site. This graph is the square of the $m$-cycle: cyclic distances one and two share a $C_i$ site. Its induced graph on $S_A$ remains connected after deleting any two units. With no deletion, cycle edges connect it. With one deletion or two adjacent deletions, the survivors form a path of distance-one edges. With two nonadjacent deletions $p, q$, distance-one edges form two paths, and the distance-two edge $\{p-1, p+1\}$ across $p$ joins them; this includes the case where a path is a single vertex. Thus shared sites force all $\gamma_j$ equal. Each $C_i$ and $Q_i$ site contains three distinct units, so it meets $S_A$ even after two deletions. All $2m$ gains equal one scalar. Conversely $RHA = 0$ puts every constant vector in the kernel. Therefore

$$\ker \mathcal{K}_A = \operatorname{span}\{\mathbf{1}_{2m}\}, \qquad \operatorname{rank} \mathcal{K}_A = 2m - 1. \tag{13}$$

The calculation uses neither the merged column nor a sign assumption on $c$. The rank is that of the linear map from $\mathbb{R}^{2m}$ to $\mathbb{R}^{m \times (m-1)}$, not of its evaluation at a particular $c$. In the notation of Li et al. (2017) (Algorithm 1 and Proposition 3.6 of the arXiv preprint), the dictionary is $H$, the data matrix is $HA$, and $R$ is their $A_\perp^*$ up to a row-basis change. Equation (13) certifies their existing rank test for every legal assignment on this structured incidence.

**Loading conditions and positive weights.** The disjoint nonempty mode supports make $A$ full column rank: one representative active row per mode forces every coefficient in a column relation to vanish. Since $H$ is injective, $HA$ has rank $m-1$. A merged pair can collapse at most two of a site's three distinct units into one mode; a zero unit leaves two active singleton modes. Thus every loading row is nonzero and mixes at least two modes. If the row matroid of $HA$ split into two nonempty direct-sum parts, a right-acting operator $\Gamma$ could have eigenvalues 1 and 2 on their spans, giving $HA\Gamma = \operatorname{diag}(q)HA$ for a nonconstant $q$. Annihilating by $R$ would put $q$ in (13), a contradiction.

Write true and candidate loadings as $F = SHU$ and $\widehat{F} = \widehat{S}H\widehat{U}$, where $U = WA$, $\widehat{U} = \widehat{W}\widehat{A}$, and $S, \widehat{S}$ are positive diagonal. In (4), $S = \operatorname{diag}(Hw)^{-1}$ and analogously for the candidate. Positive unit weights leave $U$ and $\widehat{U}$ legal, with the same supports and positive effective amplitudes. Thus the universal gate applies to $HU$, and the preceding rank and matroid arguments apply to it as well; positive site row scalings preserve these loading properties. If acting directly on the normalized $F$, its annihilator is $RS^{-1}$, not $R$. Indeed $\ker[c \mapsto R \operatorname{diag}(c)F] = \operatorname{span}(S^{-1}\mathbf{1})$; the scalar kernel (13) is asserted for $HU$.

**Affine bridge and complete comparison.** Theorem 3 gives $\operatorname{diag}(c)\widehat{F} = FJ$ with $c_v \neq 0$ and invertible $J$. Define $\xi_v = s_v/(c_v \widehat{s}_v)$, where $s_v, \widehat{s}_v$ are diagonals of $S, \widehat{S}$. Rearrange before applying the true-assignment gate:

$$H\widehat{U} = \operatorname{diag}(\xi)HUJ, \qquad R \operatorname{diag}(\xi)HU = 0.$$

The second equality follows by right-multiplying by $J^{-1}$ and using $RH = 0$. The gate forces $\xi = \kappa\mathbf{1}$, $\kappa \neq 0$, and injectivity of $H$ gives $\widehat{U} = \kappa UJ$. Every true mode has a representative row $\alpha e_k^\top$ in $U$, even if another unit is zero. Its corresponding legal row in $\widehat{U}$ is $\kappa\alpha e_k^\top J$ and has at most one nonzero entry. Every row of $J$ is therefore one-sparse; invertibility makes its occupied columns distinct. Thus $J$ is monomial. The equality also preserves zero rows and proportionality of nonzero unit rows, so an equivalent model has the same labeled zero unit or the same merged pair. It does not separate unit weights from assignment amplitudes. The latent-law conclusion follows from Theorem 3 and the common clean law after deconvolution. Candidate-specific full noise laws are outside the theorem.

# D Implementation details, configurations and seeds

## D.1 Table conventions

Every summary table uses one convention unless its caption says otherwise: bold marks the best non-oracle method for each metric, within a column when methods are rows and within a row when methods are columns. It marks the best value whichever of our variants attains it. Per-draw tables carry no bold, except Table 12, where bold marks the learned maps. Tables of differences make

no such comparison. The calibration table (Table 5) compares coefficients rather than methods and says so in its caption.

### D.2 Objective

The implemented stage-two objective is a tempered variational loss, averaged over training windows, with $X_t = X^{(t)}$ and $Z_t = Z^{(t)}$. With $n_w = N(T-\tau)$, $\bar{G}^{(0)} \in [0,1]^{D\times D}$ the instantaneous edge probabilities and $h = \operatorname{tr}\exp(\bar{G}^{(0)} \circ \bar{G}^{(0)}) - D$ the acyclicity function of Zheng et al. (2018), it is

$$\mathcal{J} = \mathbb{E}_q\big[-\log p(X_t \mid Z_t, F)\big] + \beta\big(\mathbb{E}_q[-\log p(Z_t \mid Z_{t-\tau:t-1}, G)] - \mathsf{H}(q(Z_t \mid X_t))\big) + \frac{\mathrm{KL}_F + \lambda\mathbb{E}_q|G| - \mathsf{H}(q(G))}{n_w} + \alpha_{\mathrm{AL}} h + \frac{\rho_{\mathrm{AL}}}{2} h^2. \tag{14}$$

Here $\mathcal{J}$ is the objective $\mathcal{L}_\lambda$ of Section 6.2, and $\alpha_{\mathrm{AL}}$ and $\rho_{\mathrm{AL}}$ are the multiplier and the penalty weight of the augmented Lagrangian. The reconstruction loss is the sum of squared residuals (Gaussian variance $1/2$, constants omitted). The structural term sums node-wise Gaussian conditional losses using sampled contemporaneous latents, and its interpretation as a likelihood of a directed acyclic graph applies only to acyclic masks. The weight $\beta$ multiplies the latent prior and entropy terms, not a graph KL. $\mathrm{KL}_F$ includes the categorical assignments and the stochastic log smoothing scales, and there is no amplitude KL. Relaxed factors, intermediate cyclic masks and $\beta \neq 1$ mean that this is not an unchanged evidence lower bound. Stage one, which opens the representation stage of Section 6.2, substitutes a standard normal prior and omits the graph terms. The stages on fixed latents, graph relearning and refinement, omit reconstruction, latent entropy and the factor KL.

**Observation variance.** Assumption (A0) concerns the generating class. The reconstruction term of the estimator is a Gaussian log-likelihood with the fixed variance $1/2$ on the observations centered per site and scaled by their global standard deviation, in every experiment. It weights reconstruction against the prior terms and does not estimate the noise level of the data, whose generating standard deviation is stated with each experiment.

**Acyclicity of the learned graphs.** On the fresh draws of the grid benchmark (Appendix F), $h < 4 \times 10^{-6}$ at the end of training in every run of every method trained with our graph learner, and on the committee hypergraphs and anchored structures $h \le 4.5 \times 10^{-4}$. A directed cycle of length $k$ whose edges all have probability above $1/2$ contributes $k$ closed walks of weight above $4^{-k}$ to $\operatorname{tr}\big((\bar{G}^{(0)} \circ \bar{G}^{(0)})^k\big)$, hence at least $4^{-k}/(k-1)!$ to $h$, because every term of the series is nonnegative. The thresholded instantaneous graphs therefore contain no directed cycle of length at most five on the grid benchmark and at most four on the committee hypergraphs.

### D.3 Architecture

The encoder maps the site vector of one time step, $X_t \in \mathbb{R}^L$, to the mean and log-variance of $q(Z_t \mid X_t)$ with a multilayer perceptron of two hidden layers of width 128 and LeakyReLU activations. The decoder is $\hat{X}_t = g(\hat{F} Z_t)$ with the factor family's $\hat{F}$. Our factor rescales every column of $\bar{H}A$ to unit maximum, a positive column scaling that $A$ absorbs, and the smoothed factor clips negative values at zero before this rescaling. Pair atoms carry the weights $\sum_{e \ni u,v} w_e/\delta_e$, the off-diagonal entries of $A_{\mathrm{ce}}$. The observation map $g$ of our model is the identity in every experiment except Appendix F.5, where it is the learned monotone map defined there. The heat kernel family HeatKernel sets $F_{\cdot d} \propto e^{-\gamma_d \Delta} o_d$, where $o_d$ is a relaxed one-hot center over sites (Gumbel-softmax in training, arg max at evaluation) with a uniform categorical prior and $\gamma_d > 0$ is a learned scale. Single-parent decoding SingleParent sets $F_v^\top = \operatorname{softplus}(b_v)\, \mathbf{e}_{s(v)}^\top$ with one categorical parent $s(v)$ per site, sampled in the same way. The nonlinear structural model follows Rhino (Gong et al., 2023). Every latent at every lag carries a learned 32-dimensional embedding, and a perceptron with two hidden layers of width 32 encodes each (value, embedding) pair. The encodings of the parents admitted by the graph mask are summed, and a second perceptron of the same shape maps that sum, with the child's embedding, to the conditional mean. With the heteroscedastic option a third perceptron of the same shape maps the aggregate over lagged parents only to a log-variance offset clipped to $[-6, 4]$, a simple stand-in for Rhino's conditional spline flow. The linear structural model is a masked linear map with one coefficient per lag and ordered pair. The graph holds one Bernoulli logit per lag and ordered pair, sampled with the hard Gumbel-softmax during training.

### D.4 Optimization settings

Table 4 lists the hyperparameters except the sparsity coefficient $\lambda$, which is calibrated per setting (Appendix E) and given in Section 6.2. The values are those of the grid benchmark and are shared by the factor families of Tables 1 and 2; Appendix F.2 gives the configuration of the committee and anchored experiments. All stages use Adam with gradient-norm clipping at 10, and the supplementary code gives the remaining constants.

Table 4: Optimization settings of our recipe on the grid benchmark, shared by the factor families of Tables 1 and 2; the committee and anchored experiments use the linear-dynamics configuration of Appendix F.2. The default prior weight $\beta = D/4$ of the released implementation of Wang et al. (2025) is used in graph relearning and refinement. Graph steps is the budget of stage two and of graph relearning, which then continues for the refinement steps on the same fixed latents. The coordinate kernel places each mode's center and per-axis widths in the two grid coordinates of the sites, rescaled to $[0, 1]$. Every family is initialized from the Varimax-rotated principal component loadings $\Phi$ of the observations. RBF centers and widths are the mean and spread of the site coordinates weighted by $\Phi_{vd}^2$. HeatKernel center logits and SingleParent assignment logits are proportional to $|\Phi_{vd}|$. Atom assignment logits are proportional to the unit-averaged $|\Phi_{vd}|$, normalized per mode.

| Quantity | Value |
|---|---|
| Learning rate, encoder and maps | $10^{-3}$ |
| Learning rate, factor | $10^{-2}$ |
| Learning rate, graph logits | $3 \times 10^{-2}$ |
| Learning rate, structural model | $10^{-2}$ |
| Batch size | 512 |
| Representation steps | 10000 |
| Graph steps | 4000 |
| Refinement steps | 4000 |
| Gumbel temperature at end | 0.2 |
| Augmented Lagrangian period | 200 steps |
| Linear warm start fraction | 0.5 |
| Prior weight $\beta$, stages one and two | 0.05 |
| Prior weight $\beta$, relearning and refinement | $D/4$ |
| Graph relearning on the frozen representation | on |
| Heteroscedastic noise head | on |
| Factor initialization | from Varimax-rotated loadings (caption) |
| Smoothing scale at initialization | 0.5 |
| Latent readout | encoder posterior mean |

### D.5 Seeds and how they were chosen

Two kinds of seed appear, and we keep them separate because they answer different questions.

A *model seed* controls initialization and the sampling inside training, with the data held fixed. A *data seed* controls the generation of a synthetic structure and its signals.

Model seeds were fixed as the consecutive integers starting at 0, and data seeds as the ranges stated with each experiment, before the experiment was launched. No seed was selected after the fact. The calibration of Appendix E, and every run of our model and of the comparators trained with our recipe in Appendix F, use the model seeds 0, 1 and 2; released code runs once per draw.

### D.6 Benchmark data generation and the draws

The benchmark is generated by the code released with Wang et al. (2025), at $100 \times 100$ sites, $N = T = 100$ and $\tau = 2$. The generator is not deterministic across machines, so a fixed generation seed yields different data on different hosts. Every fresh draw was therefore generated once and copied byte-identically to every host that used it (Appendix F.1), and every comparison, including

every baseline, is confined to one draw. The calibration draws of Appendix E use generation seed 0, and the fresh draws of Appendix F are labeled by their data seeds.

### D.7 BASELINE EXECUTION

The released implementation was run unmodified, for 400 epochs with 200 warm-up epochs under its own augmented Lagrangian schedule. Only missing dependencies were installed. Its default trainer configuration runs 10 epochs, fewer than the 200 epochs of its own augmented-Lagrangian warm-up, so we report the 400-epoch configuration throughout. Its visualization directory is named from a data string, a seed and a second-resolution timestamp, so two runs launched in the same second collide, and we staggered launch times.

Outside the calibration grid of Appendix E, every number we report for the two-stage pipeline (PCA-VARIMAX with our graph learner) uses its default $\beta = D/4$ and the sparsity coefficient of its setting.

## E SPARSITY CALIBRATION

The sparsity coefficient $\lambda$ is calibrated once per setting by running the graph learner on the true latents of a calibration draw, and every method we train then receives that value unless stated otherwise. Table 5 gives the grid at $D = 20$. At $D = 10$ the graph learner on the true latents of the calibration draw reaches F1 $0.666$, $0.739$, $0.843$, $0.857$, $0.868$ and $0.879$ at $\lambda = 300$, $150$, $75$, $40$, $20$ and $10$ (three seeds each). The chosen values, $\lambda^\star = 40$ at $D = 10$ and $\lambda^\star = 150$ at $D = 20$, are the grid values nearest to $0.4D^2$ (40 and 160); they are not the grid maxima, which lie at $\lambda = 10$ (0.879) and $\lambda = 75$ (0.762), and both lie within $0.025$ of those maxima. The calibration runs the graph learner on the true latents and involves no factor family, and the rule was fixed on the calibration draws before any fresh draw of Appendix F was generated.

Table 5: Calibration under nonlinear latent dynamics with a linear observation map at $D = 20$ on the calibration draw, three model seeds. ORACLE is the graph learner on the true latents, with edge F1 over all candidates and over the lagged candidates only; PCA-VARIMAX is the two-stage pipeline at the same coefficient. The graph stage runs for 8000 steps. Bold marks the protocol value $\lambda = 150$, not the maximum. The two-stage entry at $\lambda = 1200$ is taken from an earlier run (4000 graph steps).

| $\lambda$ | 1200 ($= 3D^2$) | 600 | 300 | 150 | 75 | 20 |
|---|---|---|---|---|---|---|
| ORACLE F1 | 0.492 | 0.585 | 0.667 | **0.755** | 0.762 | 0.713 |
| ORACLE lagged F1 | 0.619 | 0.730 | 0.802 | **0.886** | 0.903 | 0.844 |
| PCA-VARIMAX F1 | 0.412 | – | – | **0.528** | 0.525 | – |

The default $3D^2$ is too strong by roughly an order of magnitude on this benchmark. It was set on our own synthetic linear data at $D = 10$ and does not transfer to this benchmark. The two-stage pipeline reaches $0.528$ at the calibrated coefficient against $0.412$ at the default in an earlier run with 4000 graph steps, so the bar for this setting rises with it. Within a setting the calibrated value is the same for every method we train.

## F RECORDS OF THE CONFIRMATION EXPERIMENTS

This appendix gives the records of the confirmation experiments that Sections 6.2 and 6.3 summarize: the design, the settings and endpoints reported here, and per-draw values for the main comparisons. All numbers are copied from the analysis outputs of each experiment and rounded to three decimals unless a table states four. Every $p$-value is exact and two-sided. A draw is *won* when its difference is strictly positive. A draw with an exactly zero difference counts as not won and stays in $n$ for the sign test. The Wilcoxon signed-rank test drops zero differences and gives tied magnitudes mid-ranks. Within this appendix, exact ties occur only in the committee setting C20N10 of Appendix F.2 on F1 (three structures against each comparator).

### F.1 CONFIRMATION ON FRESH DRAWS

**Design.** Each setting uses the released generator on the $100 \times 100$ grid with lag 2, sequence length 100 and Erdős–Rényi graphs. The latent dynamics are nonlinear with history-dependent spline noise, the observation map is linear, and the settings are $D = 10$ (data seeds 7011 to 7020) and $D = 20$ (7021 to 7030). Every draw was generated once, fingerprinted and copied byte-identically to every host that used it. The recipe, configuration and sparsity are fixed and not re-tuned: ISF without smoothing and $\lambda = 40$ at $D = 10$, and ISF with smoothing and $\lambda = 150$ at $D = 20$. Our model, the same-recipe coordinate kernel and the two-stage pipeline run model seeds 0, 1 and 2, averaged within the draw before any statistic; the two-stage pipeline uses 8000 graph-learner steps and the same sparsity value. The released implementation runs once per draw, unmodified, for 400 epochs with 200 warm-up epochs, random seed 0 and its default configuration.

**Evaluation.** Both sides are scored by the same evaluator: F1 thresholds the posterior edge probabilities at 0.5, and AUPRC ranks them. For the released implementation the primary score is the maximum over every distinct latent permutation produced by its own test pass, with F1 and AUPRC maximized separately, which favors that baseline. Its native F1, which pools posterior samples, is reported alongside.

**Outcome.** All released runs finished with finite metrics, and all runs of ours and the two comparators finished. Table 6 gives the comparison with the released implementation, Table 7 the comparisons with the same-recipe coordinate kernel and the two-stage pipeline, and Table 8 every draw. In both settings our model outperforms the released implementation on every draw, on F1 and on AUPRC.

Table 6: Confirmation on fresh draws: ours minus the released implementation, with the draw as the unit. *Won* counts valid draws with a strictly positive difference (F1 / AUPRC), followed by the mean paired difference, the sign-test $p$ and the Wilcoxon signed-rank $p$. The Wilcoxon $p$ is exact and two-sided, drops zero differences, uses average ranks for tied absolute differences, and tests whether the paired differences are symmetric about zero.

| Setting | Ours | Valid | Won | Mean difference | Sign $p$ | Wilcoxon $p$ |
|---|---|---|---|---|---|---|
| nonlinear, $D = 10$ | ISF, $\lambda = 40$ | 10 | 10 / 10 | $+0.310$ / $+0.324$ | 0.002 / 0.002 | 0.002 / 0.002 |
| nonlinear, $D = 20$ | ISF with smoothing, $\lambda = 150$ | 10 | 10 / 10 | $+0.325$ / $+0.374$ | 0.002 / 0.002 | 0.002 / 0.002 |

Table 7: Confirmation on fresh draws: ours minus the same-recipe coordinate kernel and ours minus the two-stage pipeline, ten draws per setting. Entries are F1 / AUPRC. *Won* counts draws with a strictly positive difference. At $D = 20$ one of the wins against the coordinate kernel, draw 7023, is by 0.0005 F1 (0.5632 against 0.5627).

| Setting | Comparator | Won | Mean difference | Sign $p$ | Wilcoxon $p$ |
|---|---|---|---|---|---|
| nonlinear, $D = 10$ | RBF (same recipe) | 9 / 8 | $+0.058$ / $+0.088$ | 0.021 / 0.109 | 0.014 / 0.010 |
| nonlinear, $D = 10$ | PCA-VARIMAX (two-stage) | 8 / 8 | $+0.065$ / $+0.045$ | 0.109 / 0.109 | 0.020 / 0.131 |
| nonlinear, $D = 20$ | RBF (same recipe) | 9 / 10 | $+0.060$ / $+0.159$ | 0.021 / 0.002 | 0.010 / 0.002 |
| nonlinear, $D = 20$ | PCA-VARIMAX (two-stage) | 10 / 10 | $+0.181$ / $+0.186$ | 0.002 / 0.002 | 0.002 / 0.002 |

**What the comparison contains.** The comparison with the released implementation is our full model against that implementation, so likelihood, noise model and schedule differ as well as the spatial factor. The comparisons with the same-recipe coordinate kernel vary only the factor family. Our model trains on all 100 sequences of a draw and the released implementation on its training split of 80; Appendix F.9 repeats our side on the same 80 sequences. The released implementation is scored at its most favorable latent permutation.

### F.2 UNIT-LEVEL AND PAIR-LEVEL ATOMS ON COMMITTEE HYPERGRAPHS

**Premise and design.** The generator's spatial factors are hyperedge-atom footprints, so the data follow the membership relation, and the rewire control keeps this fixed. The latent dimension is $D =$

Table 8: Confirmation on fresh draws, every draw. Ours is the mean over three model seeds of the fixed configuration of Table 6. *Released, primary* is the released implementation scored with our evaluator at the latent permutation most favorable to it, per metric; *native F1* is its own F1, which pools posterior samples. Draw labels give the setting and the data seed.

| | Ours | | Released, primary | | Released |
|---|---|---|---|---|---|
| Draw | F1 | AUPRC | F1 | AUPRC | native F1 |
| NL10 7011 | 0.752 | 0.822 | 0.270 | 0.241 | 0.206 |
| NL10 7012 | 0.720 | 0.685 | 0.293 | 0.239 | 0.199 |
| NL10 7013 | 0.802 | 0.805 | 0.490 | 0.483 | 0.404 |
| NL10 7014 | 0.803 | 0.800 | 0.412 | 0.396 | 0.248 |
| NL10 7015 | 0.585 | 0.624 | 0.148 | 0.170 | 0.082 |
| NL10 7016 | 0.605 | 0.629 | 0.356 | 0.371 | 0.221 |
| NL10 7017 | 0.562 | 0.651 | 0.537 | 0.613 | 0.416 |
| NL10 7018 | 0.754 | 0.725 | 0.343 | 0.326 | 0.211 |
| NL10 7019 | 0.416 | 0.365 | 0.200 | 0.211 | 0.134 |
| NL10 7020 | 0.690 | 0.700 | 0.542 | 0.514 | 0.467 |
| NL20 7021 | 0.774 | 0.745 | 0.302 | 0.219 | 0.171 |
| NL20 7022 | 0.659 | 0.614 | 0.292 | 0.301 | 0.228 |
| NL20 7023 | 0.563 | 0.639 | 0.340 | 0.320 | 0.252 |
| NL20 7024 | 0.676 | 0.686 | 0.283 | 0.186 | 0.196 |
| NL20 7025 | 0.647 | 0.691 | 0.435 | 0.361 | 0.356 |
| NL20 7026 | 0.606 | 0.564 | 0.405 | 0.338 | 0.290 |
| NL20 7027 | 0.602 | 0.625 | 0.386 | 0.310 | 0.291 |
| NL20 7028 | 0.547 | 0.600 | 0.271 | 0.244 | 0.205 |
| NL20 7029 | 0.714 | 0.671 | 0.276 | 0.274 | 0.165 |
| NL20 7030 | 0.723 | 0.696 | 0.272 | 0.233 | 0.186 |

10 with an identity observation map, and the latent dynamics are linear with lag 2 and $T = 100$. The latent graph has $D$ instantaneous edges forming a random acyclic graph, $D$ random edges at each lag and a self-loop on every latent at lag one. Edge weights are uniform on $[0.2, 0.6]$ with a random sign and are redrawn until the reduced-form companion matrix has spectral radius at most $0.95$. Innovations are $\mathcal{N}(0, 0.5^2)$, the first 100 steps are discarded, and the latents are standardized. Each committee hyperedge is a uniformly random site subset of uniformly random size, an uncovered site joins a random hyperedge, and a structure is redrawn until it is connected. Each latent receives six random hyperedges with amplitudes uniform on $(0.6, 1.4)$, and every column of $F$ is scaled to unit maximum. The estimator uses its linear-dynamics configuration: a linear structural model, 1200 representation steps, prior weight $\beta = D/4$ in every stage, $\lambda = 3D^2 = 300$, no linear warm start, no graph relearning and no heteroscedastic head. Learning rates, batch size, graph and refinement steps and the Gumbel temperature are those of Table 4. The setting C20 is the committee hypergraph of Section 6.3 ($L = 300$, $M = 60$, hyperedges of 20 to 40 sites, six hyperedges per latent) with additive Gaussian observation noise of standard deviation 2.0 and $N = 100$. C15 is the same at noise 1.5, and C20N10 is C20 with $N = 10$. Each setting has ten structures with seeds 8101 to 8110; the seed fixes both the structure and the data, so the same ten fresh structures are used in every setting. The methods are ISF without smoothing on the true memberships, on their clique expansion (pair atoms) and on rewired memberships, and the heat kernel on the same pairwise operator (HEATKERNEL). The rewiring makes four accepted degree- and size-preserving bipartite swaps per incidence entry, and every rewired run was checked to have completed all its swaps. Model seeds 0, 1 and 2 are averaged within each structure. The primary endpoint is latent MCC, with the structure as the unit; factor cosine, F1 and AUPRC are secondary.

**Outcome.** All runs finished with finite MCC, and every rewired run completed its swaps. Table 9 gives the settings, contrasts and metrics. In C20 and C15, ISF on the true memberships leads each of the three comparators on MCC in ten of ten structures. In C20N10, with ten sequences per structure, hyperedge atoms lead all three comparators on MCC and factor cosine in ten of ten structures.

### F.3 UNITS RECONSTRUCTED FROM THE WEIGHTED CLIQUE EXPANSION

**Design.** The input is the weighted clique expansion $A_{\mathrm{ce}} = HW\mathrm{diag}(\delta)^{-1}H^{\top}$ with its diagonal, the operator of Proposition 1 and Theorem 2. Its row sums are the site degrees, so the heat kernel is a function of $A_{\mathrm{ce}}$, and the atoms and weights of ISF on the clique expansion are its off-diagonal

Table 9: Committee hypergraphs: ISF on the true memberships minus each comparator, ten structures per setting. Each entry gives the structures won out of ten (strictly positive difference), the mean difference, the sign-test $p$ and the Wilcoxon signed-rank $p$. In C20N10 three F1 structures per comparator are exact ties; they count as not won and stay in $n = 10$ for the sign test, and the Wilcoxon test drops them. The factor cosine is $D^{-1} \sum_d |\langle F_{\cdot d}, \hat{F}_{\cdot \hat{\pi}(d)} \rangle| / (\|F_{\cdot d}\| \, \|\hat{F}_{\cdot \hat{\pi}(d)}\|)$ with the permutation $\hat{\pi}$ of Section 6.1.

| Setting | vs | MCC | Factor cosine | F1 | AUPRC |
|---|---|---|---|---|---|
| C20 | pair atoms | 10, +0.089, 0.002, 0.002 | 10, +0.114, 0.002, 0.002 | 10, +0.126, 0.002, 0.002 | 9, +0.104, 0.021, 0.004 |
| | HEATKERNEL | 10, +0.327, 0.002, 0.002 | 10, +0.496, 0.002, 0.002 | 10, +0.438, 0.002, 0.002 | 10, +0.331, 0.002, 0.002 |
| | rewired | 10, +0.443, 0.002, 0.002 | 10, +0.551, 0.002, 0.002 | 10, +0.454, 0.002, 0.002 | 10, +0.342, 0.002, 0.002 |
| C15 | pair atoms | 10, +0.044, 0.002, 0.002 | 8, +0.041, 0.109, 0.010 | 9, +0.049, 0.021, 0.064 | 8, +0.036, 0.109, 0.084 |
| | HEATKERNEL | 10, +0.313, 0.002, 0.002 | 10, +0.514, 0.002, 0.002 | 10, +0.546, 0.002, 0.002 | 10, +0.484, 0.002, 0.002 |
| | rewired | 10, +0.464, 0.002, 0.002 | 10, +0.592, 0.002, 0.002 | 10, +0.584, 0.002, 0.002 | 10, +0.513, 0.002, 0.002 |
| C20N10 | pair atoms | 10, +0.127, 0.002, 0.002 | 10, +0.202, 0.002, 0.002 | 6, +0.021, 0.754, 0.156 | 8, +0.029, 0.109, 0.037 |
| | HEATKERNEL | 10, +0.291, 0.002, 0.002 | 10, +0.445, 0.002, 0.002 | 7, +0.059, 0.344, 0.016 | 9, +0.075, 0.021, 0.004 |
| | rewired | 10, +0.371, 0.002, 0.002 | 10, +0.470, 0.002, 0.002 | 7, +0.052, 0.344, 0.016 | 10, +0.084, 0.002, 0.002 |

support and entries. Since $A_{\mathrm{ce}} = BB^\top$ with $B = H\mathrm{diag}(w/\delta)^{1/2} \geq 0$, we set $k = \mathrm{rank}(A_{\mathrm{ce}})$, the number of singular values above $10^{-8}$ times the largest, which equals $M = 60$ on these structures. We fit $B \in \mathbb{R}^{L \times k}_{\geq 0}$ by the multiplicative update of Ding et al. (2005, eq. 34) with its step parameter equal to $1/2$,

$$B \leftarrow B \circ \Big( \tfrac{1}{2} + \tfrac{1}{2} \, (A_{\mathrm{ce}} B) \oslash (BB^\top B) \Big),$$

with $\circ$ and $\oslash$ the entrywise product and division, for 20000 steps from four random starts, and keep the start with the smallest $\|A_{\mathrm{ce}} - BB^\top\|_F / \|A_{\mathrm{ce}}\|_F$. Unit $j$ is $\hat{e}_j = \{v : B_{vj} > \tfrac{1}{2} \max_u B_{uj}\}$ with weight $\hat{w}_j = |\hat{e}_j| \cdot \mathrm{mean}_{v \in \hat{e}_j} B^2_{vj}$, and a site outside every unit joins the unit with the largest entry of $B_{v\cdot}$. No other information enters. Data, structures, configuration and seeds are those of Appendix F.2 (C20 and C15). ISF on the reconstructed units, ISF on the true memberships and ISF on the clique expansion are rerun together with model seeds 0, 1 and 2.

**Outcome.** All 180 runs finished. In each of the 60 runs on reconstructed units the reconstruction has relative error below $10^{-4}$, no site is left outside a unit, and all 60 hyperedges are recovered exactly. Hyperedge atoms on the reconstructed units have a higher MCC than pair atoms on 10 of 10 structures at noise 2.0 and 9 of 10 at noise 1.5 (sign test $p = 0.002$ and 0.021). Against hyperedge atoms on the true memberships they are higher on 7 and 6 of 10 structures, with mean differences of $+0.008$ and $+0.0001$ (Table 10). Because the reconstructed units equal the true ones, the two differ only in the order of the atoms and in weights that agree up to the reconstruction error. The order of the atoms changes the random samples each atom receives. The reruns agree with Table 3 within 0.003 in mean MCC, 0.006 in mean F1 and 0.02 in mean AUPRC. In AUPRC, the reconstructed units exceed pair atoms on 10 and 9 of 10 structures. Against the heat kernel and the rewired memberships of Appendix F.2, which use the same data and seeds, the reconstructed units have a higher MCC on every structure at both noise levels.

### F.4 ANCHORED COMMITTEE STRUCTURES

**Design.** Committee hypergraphs with $L = 450$, $M = 60$, hyperedges of 20 to 40 sites, hyperedge-atom ground truth, $D = 10$, six hyperedges per latent, an identity observation map, $N = 100$ and noise 0.5, with the generator required to give every latent an anchor site. Seeds 8201 to 8210 fix structure and data. For every structure and seed we record the anchored fraction, which must be 1 for (C1), bridge connectivity for (C2), and the multi-parent fraction. The estimator is unsmoothed ISF on the given hyperedges, the class of Theorem 1, in the configuration of Appendix F.2, with model seeds 0, 1 and 2; the endpoint is the latent MCC averaged over them.

**Outcome.** All runs finished. (C1) and (C2) hold on all ten structures, and between 87.8% and 93.3% of the sites are multi-parent (Table 11). The latent MCC lies between 0.9866 and 0.9887 on nine of the ten structures.

Table 10: ISF on units reconstructed from the weighted clique expansion: latent MCC and edge AUPRC per structure, mean over three model seeds, four decimals. The columns are ISF on the clique expansion (pair atoms), on the true memberships and on the reconstructed units. The table carries no bold.

| | noise 2.0 | | | noise 1.5 | | |
|---|---|---|---|---|---|---|
| Structure | Pair atoms | True (ours) | Reconstructed (ours) | Pair atoms | True (ours) | Reconstructed (ours) |
| *Latent MCC* | | | | | | |
| 8101 | 0.7535 | 0.7755 | 0.8154 | 0.8524 | 0.8668 | 0.9048 |
| 8102 | 0.5767 | 0.7382 | 0.7516 | 0.7545 | 0.8377 | 0.8071 |
| 8103 | 0.6373 | 0.7719 | 0.7723 | 0.8002 | 0.8481 | 0.8530 |
| 8104 | 0.7192 | 0.7859 | 0.8039 | 0.8676 | 0.9248 | 0.9247 |
| 8105 | 0.6766 | 0.7883 | 0.7888 | 0.8833 | 0.8904 | 0.8567 |
| 8106 | 0.8268 | 0.8490 | 0.8487 | 0.9116 | 0.9173 | 0.9174 |
| 8107 | 0.6167 | 0.7785 | 0.7875 | 0.7737 | 0.8612 | 0.8802 |
| 8108 | 0.8034 | 0.8316 | 0.8506 | 0.9085 | 0.9119 | 0.9118 |
| 8109 | 0.6736 | 0.7758 | 0.7663 | 0.7800 | 0.8623 | 0.8659 |
| 8110 | 0.7379 | 0.8353 | 0.8276 | 0.8581 | 0.8962 | 0.8963 |
| Mean | 0.7022 | 0.7930 | 0.8013 | 0.8390 | 0.8817 | 0.8818 |
| *Edge AUPRC* | | | | | | |
| 8101 | 0.4411 | 0.5537 | 0.6041 | 0.6642 | 0.7139 | 0.7098 |
| 8102 | 0.3995 | 0.3761 | 0.4132 | 0.5677 | 0.4410 | 0.4687 |
| 8103 | 0.3141 | 0.4492 | 0.4515 | 0.5457 | 0.5999 | 0.5833 |
| 8104 | 0.4033 | 0.5744 | 0.5331 | 0.5778 | 0.6557 | 0.6610 |
| 8105 | 0.2120 | 0.3790 | 0.3593 | 0.6091 | 0.6212 | 0.6129 |
| 8106 | 0.5920 | 0.7909 | 0.7294 | 0.9579 | 0.9554 | 0.9590 |
| 8107 | 0.3037 | 0.3965 | 0.3667 | 0.3892 | 0.4593 | 0.4772 |
| 8108 | 0.5650 | 0.6745 | 0.7489 | 0.8961 | 0.9228 | 0.9247 |
| 8109 | 0.2982 | 0.3734 | 0.3180 | 0.4814 | 0.5870 | 0.5692 |
| 8110 | 0.3445 | 0.5445 | 0.4480 | 0.5695 | 0.7253 | 0.7251 |
| Mean | 0.3873 | 0.5112 | 0.4972 | 0.6259 | 0.6682 | 0.6691 |

Table 11: Anchored committee structures, noise $0.5$. The ISF MCC is the mean over three model seeds, to four decimals; the per-seed values are listed for structure 8201, and for 8202 all three equal $0.987$. (C1) holds when the anchored fraction is $1$, and (C2) when the mode co-occurrence graph is connected. The last column is the fraction of multi-parent sites.

| Structure | ISF MCC | ISF seeds | (C1) | (C2) | Multi-parent |
|---|---|---|---|---|---|
| 8201 | 0.8625 | 0.871 / 0.880 / 0.837 | yes | yes | 0.933 |
| 8202 | 0.9870 | 0.987 (all three) | yes | yes | 0.913 |
| 8203 | 0.9885 | | yes | yes | 0.898 |
| 8204 | 0.9869 | | yes | yes | 0.922 |
| 8205 | 0.9866 | | yes | yes | 0.891 |
| 8206 | 0.9887 | | yes | yes | 0.878 |
| 8207 | 0.9879 | | yes | yes | 0.902 |
| 8208 | 0.9878 | | yes | yes | 0.898 |
| 8209 | 0.9883 | | yes | yes | 0.918 |
| 8210 | 0.9874 | | yes | yes | 0.896 |

### F.5 Anchored structures with monotone observation maps

**Design.** The structures, footprints and seeds are those of Appendix F.4 (seeds 8201 to 8210 fix structure and data; the structure hashes coincide). Every site is now observed through its own invertible map, $X_v^{(t)} = g_v\big(F_v^\top Z^{(t)}\big) + \varepsilon_v^{(t)}$, with $\varepsilon_v^{(t)} \sim \mathcal{N}(0, 0.5^2)$ added after the map and

$$g_v(y) = y + \mu_v \tanh(\nu_v y + \zeta_v) - \mu_v \tanh(\zeta_v),$$
$$\nu_v \sim \mathcal{U}(0.5, 2), \qquad \mu_v \sim \mathcal{U}(-0.9, 0.9)/\nu_v, \qquad \zeta_v \sim \mathcal{U}(-0.5, 0.5),$$

drawn independently per site. Then $g_v' = 1 + \mu_v\nu_v \operatorname{sech}^2(\nu_v y + \zeta_v) \geq 1 - |\mu_v\nu_v| > 0.1$, so every $g_v$ is a $C^2$ diffeomorphism of $\mathbb{R}$, as in Assumption (A1). The estimator is unsmoothed ISF, the class of Theorem 1, with a learned map of the same form per site, $\hat{g}_v(y) = y + \hat{\mu}_v \tanh(\hat{\nu}_v y + \hat{\zeta}_v) - \hat{\mu}_v \tanh(\hat{\zeta}_v)$. A smooth reparameterization enforces $\hat{\nu}_v \geq 0.25$ and $|\hat{\mu}_v\hat{\nu}_v| < 0.95$. The maps are initialized at the identity and trained jointly with the factor, the encoder and the graph in the configuration of Appendix F.2. The family is known and its $3L$ parameters are not. The reference is the same estimator with the identity map. Model seeds are 0, 1 and 2; the endpoint is the latent MCC averaged over them.

**Outcome.** All runs finished. (C1) and (C2) hold on all ten structures. With the learned maps, the latent MCC is at least $0.95$ on every structure and exceeds the MCC of the identity map on every structure (sign test $p = 0.002$). The reconstruction error is lower on every structure as well (Table 12).

Table 12: Anchored structures with site-specific monotone observation maps, noise $0.5$. Latent MCC (mean over three model seeds, four decimals) and reconstruction mean squared error of unsmoothed ISF with the identity map and with learned monotone maps. The error is computed on the observations centered per site and scaled by their global standard deviation. Per-seed MCC with learned maps: 8201 0.989 / 0.885 / 0.989; 8205 0.987 / 0.917 / 0.986. The last three columns are as in Table 11.

| | Identity map | | Learned monotone maps (ours) | | | | |
|---|---|---|---|---|---|---|---|
| Structure | MCC | Recon. MSE | MCC | Recon. MSE | (C1) | (C2) | Multi-parent |
| 8201 | 0.9504 | 0.475 | **0.9540** | **0.403** | yes | yes | 0.933 |
| 8202 | 0.9858 | 0.505 | **0.9879** | **0.444** | yes | yes | 0.913 |
| 8203 | 0.9880 | 0.478 | **0.9896** | **0.421** | yes | yes | 0.898 |
| 8204 | 0.9855 | 0.488 | **0.9878** | **0.425** | yes | yes | 0.922 |
| 8205 | 0.9435 | 0.526 | **0.9632** | **0.464** | yes | yes | 0.891 |
| 8206 | 0.9878 | 0.481 | **0.9895** | **0.423** | yes | yes | 0.878 |
| 8207 | 0.9870 | 0.525 | **0.9888** | **0.463** | yes | yes | 0.902 |
| 8208 | 0.9868 | 0.505 | **0.9886** | **0.444** | yes | yes | 0.898 |
| 8209 | 0.9886 | 0.478 | **0.9898** | **0.416** | yes | yes | 0.918 |
| 8210 | 0.9864 | 0.502 | **0.9883** | **0.443** | yes | yes | 0.896 |
| Mean | 0.9790 | 0.496 | **0.9828** | **0.435** | | | |

### F.6 Released implementation with a calibrated sparsity coefficient

**Motivation.** Our sparsity coefficient $\lambda$ was chosen with generator ground truth (Appendix E), while the confirmation of Appendix F.1 ran the released implementation at its default sparsity coefficient. To give the released implementation the same kind of access, we calibrated its coefficient `graph_sparsity_factor`, denoted $k$ here, on separate draws. It enters the released objective as $k \sum_{ij} G_{ij}/n_{\text{frag}}$, with $n_{\text{frag}}$ the number of training fragments, the same normalization as our term $\lambda\, \mathbb{E}_q|G|/n_w$ over $n_w$ training windows (Appendix D.2). Its default is $k = 10$; our values are $\lambda = 40$ and $150$ in the two settings. The access differs in kind: our $\lambda$ was chosen with the true latents on earlier draws, and the released coefficient is chosen with the true graph on a separate draw.

**Calibration.** We use one calibration draw per setting of Appendix F.1, with the same generator settings and not used in that confirmation's analysis. The draws are data seed 7001 at $D = 10$

(a pipeline-test draw generated with the confirmation draws and excluded from their analysis) and 7042 at $D = 20$. The released implementation runs unmodified apart from $k$, for 400 epochs with 200 warm-up epochs and random seed 0, over the grid $k \in \{3, 10, 30, 100, 300\}$, which brackets its default and our values. Every run is scored as the primary comparison of Appendix F.1: F1 at posterior probability above $0.5$ and AUPRC, each at the most favorable of the latent alignments produced by the released test pass.

**Selection rule.** Per setting, $k^*$ maximizes the calibration-draw F1; ties go to the default 10, then to the smaller $k$. If $k^* = 10$, the confirmation values of Appendix F.1 are already the calibrated ones and the setting is not rerun. Otherwise the released implementation is rerun with $k^*$ on the ten confirmation draws of that setting, with random seed 0 and the same scoring; our values and those of the comparators are unchanged. All calibration runs finished. Table 13 gives every value. The selected coefficients are $k^* = 30$ at $D = 10$ and $k^* = 3$ at $D = 20$.

Table 13: Calibration of the released implementation's sparsity coefficient $k$ on one separate draw per setting, F1 / AUPRC at the most favorable latent alignment. Bold marks the F1 maximum, which selects $k^*$; the default is $k = 10$.

| Setting (draw) | $k = 3$ | $k = 10$ (default) | $k = 30$ | $k = 100$ | $k = 300$ | $k^*$ |
|---|---|---|---|---|---|---|
| nonlinear, $D = 10$ (7001) | 0.489 / 0.403 | 0.439 / 0.431 | **0.500** / 0.479 | 0.143 / 0.340 | 0.000 / 0.187 | 30 |
| nonlinear, $D = 20$ (7042) | **0.291** / 0.217 | 0.237 / 0.210 | 0.225 / 0.204 | 0.163 / 0.153 | 0.000 / 0.094 | 3 |

**Confirmation.** Table 14 gives every draw of the two reruns and Table 15 the summary. Our values and the default-coefficient values are those of Table 8.

Table 14: Released implementation at the calibrated coefficient $k^*$ on the confirmation draws, both scored at the latent alignment most favorable to it. Ours and the default-coefficient values ($k = 10$) are copied from Table 8.

| | Ours | | Released, $k = 10$ | | Released, $k^*$ | |
|---|---|---|---|---|---|---|
| Draw | F1 | AUPRC | F1 | AUPRC | F1 | AUPRC |
| NL10 7011 | 0.752 | 0.822 | 0.270 | 0.241 | 0.258 | 0.250 |
| NL10 7012 | 0.720 | 0.685 | 0.293 | 0.239 | 0.316 | 0.247 |
| NL10 7013 | 0.802 | 0.805 | 0.490 | 0.483 | 0.419 | 0.380 |
| NL10 7014 | 0.803 | 0.800 | 0.412 | 0.396 | 0.357 | 0.325 |
| NL10 7015 | 0.585 | 0.624 | 0.148 | 0.170 | 0.154 | 0.176 |
| NL10 7016 | 0.605 | 0.629 | 0.356 | 0.371 | 0.316 | 0.308 |
| NL10 7017 | 0.562 | 0.651 | 0.537 | 0.613 | 0.581 | 0.596 |
| NL10 7018 | 0.754 | 0.725 | 0.343 | 0.326 | 0.323 | 0.291 |
| NL10 7019 | 0.416 | 0.365 | 0.200 | 0.211 | 0.222 | 0.239 |
| NL10 7020 | 0.690 | 0.700 | 0.542 | 0.514 | 0.242 | 0.351 |
| NL20 7021 | 0.774 | 0.745 | 0.302 | 0.219 | 0.294 | 0.195 |
| NL20 7022 | 0.659 | 0.614 | 0.292 | 0.301 | 0.359 | 0.397 |
| NL20 7023 | 0.563 | 0.639 | 0.340 | 0.320 | 0.342 | 0.333 |
| NL20 7024 | 0.676 | 0.686 | 0.283 | 0.186 | 0.269 | 0.223 |
| NL20 7025 | 0.647 | 0.691 | 0.435 | 0.361 | 0.424 | 0.401 |
| NL20 7026 | 0.606 | 0.564 | 0.405 | 0.338 | 0.370 | 0.299 |
| NL20 7027 | 0.602 | 0.625 | 0.386 | 0.310 | 0.346 | 0.339 |
| NL20 7028 | 0.547 | 0.600 | 0.271 | 0.244 | 0.197 | 0.198 |
| NL20 7029 | 0.714 | 0.671 | 0.276 | 0.274 | 0.165 | 0.269 |
| NL20 7030 | 0.723 | 0.696 | 0.272 | 0.233 | 0.295 | 0.255 |

Table 15: Calibrated released implementation, ours minus the released implementation at $k^*$, with the draw as the unit. Entries are F1 / AUPRC.

| Setting | $k^*$ | Valid | Won | Mean difference | Sign $p$ |
|---|---|---|---|---|---|
| nonlinear, $D = 10$ | 30 | 10 | 9 / 10 | $+0.350$ / $+0.364$ | 0.021 / 0.002 |
| nonlinear, $D = 20$ | 3 | 10 | 10 / 10 | $+0.345$ / $+0.362$ | 0.002 / 0.002 |

### F.7 Released two-stage and sequence-model baselines

**Design.** Among the baselines in the public code of Wang et al. (2025), we run four with their released configurations on all forty draws of Appendices F.1 and F.11. They are Mapped-PCMCI on Varimax-rotated principal components (Tibau et al., 2022), TPCA-PCMCI on tensor principal components (Babii et al., 2025), LEAP (Yao et al., 2022b) and TDRL (Yao et al., 2022a). In the released code the graph of all four comes from an unconditional lagged-dependence test at lags $1$ to $\tau$, so these baselines return no instantaneous edges. For the two PCMCI baselines, Wang et al. (2025) describe this step as PCMCI+ (Runge, 2020). Every draw receives an isolated copy of the code with that draw's data. The four baselines run one after another with random seed $0$, one run per draw as for the released joint model, on the released split of sequences $0$ to $79$ for training and $80$ to $99$ for validation. The released evaluation function reads the maximum lag before assigning it, which stops LEAP and TDRL at test time. Each copy assigns it at the top of that function, and no other line of the code changes. Scores are the native test metrics of the released code, the edge F1 after its own latent matching; the released code reports no AUPRC for these methods. Our values are those of Appendices F.1 and F.11 (our evaluator, mean over three model seeds), and the draw is the unit. A failed run is repeated once with the same seed; a draw that still fails leaves that baseline's count.

**Scores for AUPRC.** The released code returns binary graphs only. For AUPRC, each code copy also stores the statistics of the lagged test at test time, before the released thresholding and rescaling. They are saved next to the graph that its F1 uses, and no line of the released computation changes. A lagged candidate is scored by the absolute statistic. An instantaneous candidate is not examined by the released test and is scored by zero. AUPRC is the average precision at distinct score thresholds over the candidates of Section 6.1, pooled over the test sequences as the released F1 pools them. For Mapped-PCMCI and TPCA-PCMCI, the pass with stored scores reproduces the native F1 exactly on all forty draws. LEAP and TDRL were re-evaluated from their final-epoch checkpoints. Their inference samples the latent state at every step. The second pass therefore differs from the first by at most $0.006$ in F1 on a draw and by at most $0.001$ in mean F1. Both F1 and AUPRC of these two methods come from the second pass. Our AUPRC follows the definition in Section 6.1 without tie handling. On eight reruns of our model and of the coordinate kernel that store the posterior probabilities, AUPRC with and without tie handling coincide.

**Outcome.** All $160$ runs completed and none was excluded. Our model has the higher F1 on every draw against each of the four baselines in both settings. Its AUPRC is higher on every draw except one draw against Mapped-PCMCI at $D = 10$ (Table 16).

Table 16: Released two-stage and sequence-model baselines on the forty draws of the grid benchmark, twenty per setting. The baseline column gives each baseline's mean native edge F1 and the AUPRC from its stored test statistics. The remaining columns give our means on the same draws (our evaluator), the draws on which our value is higher and the sign-test $p$, each for F1 and then AUPRC. LEAP and TDRL values come from the evaluation pass with stored scores.

| Setting | Baseline | Baseline F1 / AUPRC | Our F1 / AUPRC | Won | Sign $p$ |
|---|---|---|---|---|---|
| $D = 10$ | Mapped-PCMCI | 0.361 / 0.390 | **0.666 / 0.699** | 20/20, 19/20 | $1.9 \times 10^{-6}$, $4.0 \times 10^{-5}$ |
| | TPCA-PCMCI | 0.084 / 0.106 | **0.666 / 0.699** | 20/20, 20/20 | $1.9 \times 10^{-6}$, $1.9 \times 10^{-6}$ |
| | LEAP | 0.077 / 0.107 | **0.666 / 0.699** | 20/20, 20/20 | $1.9 \times 10^{-6}$, $1.9 \times 10^{-6}$ |
| | TDRL | 0.095 / 0.118 | **0.666 / 0.699** | 20/20, 20/20 | $1.9 \times 10^{-6}$, $1.9 \times 10^{-6}$ |
| $D = 20$ | Mapped-PCMCI | 0.290 / 0.260 | **0.657 / 0.662** | 20/20, 20/20 | $1.9 \times 10^{-6}$, $1.9 \times 10^{-6}$ |
| | TPCA-PCMCI | 0.042 / 0.050 | **0.657 / 0.662** | 20/20, 20/20 | $1.9 \times 10^{-6}$, $1.9 \times 10^{-6}$ |
| | LEAP | 0.044 / 0.051 | **0.657 / 0.662** | 20/20, 20/20 | $1.9 \times 10^{-6}$, $1.9 \times 10^{-6}$ |
| | TDRL | 0.048 / 0.053 | **0.657 / 0.662** | 20/20, 20/20 | $1.9 \times 10^{-6}$, $1.9 \times 10^{-6}$ |

### F.8 Rewired grid incidence

**Design.** This experiment asks whether the same-recipe advantage over the coordinate kernel on the grid benchmark (Appendix F.1) depends on which sites share atoms, or only on the parameterization as a dictionary of two-site atoms. The data of the ten confirmation draws per setting are unchanged:

the data fingerprint of every run equals that of the confirmation. Our factor, recipe and $\lambda$ receive the $100 \times 100$ grid incidence after degree- and size-preserving bipartite double-edge swaps, four accepted swaps per incidence entry ($158{,}400$ swaps; every run reached the target). This is a capacity-matched control: the atom count, the atom sizes, every site degree, the recipe and $\lambda$ are those of the true grid. Each draw receives its own rewired incidence, shared by its three model seeds; at $D = 20$ ISF recomputes its heat smoothing on the rewired incidence. The comparators are the confirmation's own values on the same draws and seeds.

**Endpoint.** Per setting, with the draw as the unit and a tie counted as not won, we count the draws on which the rewired factor falls below the same-recipe coordinate kernel on F1. Differences to our factor on the true grid are reported as effect sizes.

**Results.** All 60 runs finished. In both settings the rewired factor falls below the coordinate kernel on F1 on all ten draws (Table 17). At $D = 10$ ISF uses no smoothing, so the loss is due to the incidence alone. The loss arises before the graph stage: the latents are recovered with mean MCC $0.722$ and $0.685$, against $0.968$ on the true grid and $0.935$ and $0.907$ for the coordinate kernel. Choosing, per draw, the kernel seed with the lowest final training objective (the rule of multi-start selection) yields mean F1 $0.605$ and $0.604$ against $0.611$ and $0.591$ for the seed mean. Our factor's seed mean exceeds this selected kernel in 9 and 8 of 10 draws.

Table 17: Rewired grid incidence, with the draw as the unit. RW is ISF ($\lambda = 40$ at $D = 10$; with smoothing and $\lambda = 150$ at $D = 20$) on the rewired incidence, RBF the same-recipe coordinate kernel and Grid ISF on the true grid, all means over three model seeds. Differences RW $-$ Grid are negative on all ten draws in both settings for F1, AUPRC and MCC (sign test $p = 0.002$ each).

| Setting | RW loses to RBF (F1) | Mean F1 RW / RBF / Grid | RW $-$ Grid: F1 / AUPRC / MCC |
|---|---|---|---|
| nonlinear, $D = 10$ | 10 / 10 | 0.324 / 0.611 / **0.669** | $-0.344$ / $-0.432$ / $-0.246$ |
| nonlinear, $D = 20$ | 10 / 10 | 0.275 / 0.591 / **0.651** | $-0.377$ / $-0.488$ / $-0.283$ |

Table 18: Rewired grid incidence, every draw (means over three model seeds). At draw 7023 the F1 values of RBF and of our factor are $0.5627$ and $0.5632$ before rounding.

| Draw | F1 RW | F1 RBF | F1 Grid | AUPRC RW | AUPRC Grid | MCC RW | MCC Grid | MCC RBF |
|---|---|---|---|---|---|---|---|---|
| NL10 7011 | 0.437 | 0.684 | 0.752 | 0.336 | 0.822 | 0.755 | 0.996 | 0.954 |
| NL10 7012 | 0.274 | 0.628 | 0.720 | 0.230 | 0.685 | 0.631 | 0.939 | 0.915 |
| NL10 7013 | 0.359 | 0.782 | 0.802 | 0.281 | 0.805 | 0.813 | 0.996 | 0.998 |
| NL10 7014 | 0.452 | 0.676 | 0.803 | 0.330 | 0.800 | 0.815 | 0.998 | 0.904 |
| NL10 7015 | 0.310 | 0.612 | 0.585 | 0.233 | 0.624 | 0.803 | 0.996 | 0.997 |
| NL10 7016 | 0.221 | 0.600 | 0.605 | 0.167 | 0.629 | 0.680 | 0.967 | 0.898 |
| NL10 7017 | 0.279 | 0.551 | 0.562 | 0.201 | 0.651 | 0.621 | 0.854 | 0.833 |
| NL10 7018 | 0.245 | 0.659 | 0.754 | 0.205 | 0.725 | 0.710 | 0.991 | 0.990 |
| NL10 7019 | 0.243 | 0.295 | 0.416 | 0.153 | 0.365 | 0.610 | 0.953 | 0.868 |
| NL10 7020 | 0.423 | 0.625 | 0.690 | 0.348 | 0.700 | 0.781 | 0.992 | 0.989 |
| NL20 7021 | 0.300 | 0.609 | 0.774 | 0.173 | 0.745 | 0.648 | 0.995 | 0.921 |
| NL20 7022 | 0.512 | 0.681 | 0.659 | 0.363 | 0.614 | 0.836 | 0.958 | 0.947 |
| NL20 7023 | 0.266 | 0.563 | 0.563 | 0.155 | 0.639 | 0.626 | 0.952 | 0.903 |
| NL20 7024 | 0.259 | 0.658 | 0.676 | 0.160 | 0.686 | 0.777 | 0.986 | 0.943 |
| NL20 7025 | 0.297 | 0.625 | 0.647 | 0.167 | 0.691 | 0.690 | 0.951 | 0.881 |
| NL20 7026 | 0.288 | 0.543 | 0.606 | 0.142 | 0.564 | 0.685 | 0.949 | 0.904 |
| NL20 7027 | 0.224 | 0.568 | 0.602 | 0.138 | 0.625 | 0.761 | 0.958 | 0.898 |
| NL20 7028 | 0.129 | 0.477 | 0.547 | 0.078 | 0.600 | 0.550 | 0.984 | 0.841 |
| NL20 7029 | 0.205 | 0.572 | 0.714 | 0.099 | 0.671 | 0.586 | 0.953 | 0.896 |
| NL20 7030 | 0.266 | 0.617 | 0.723 | 0.176 | 0.696 | 0.696 | 0.996 | 0.934 |

### F.9 Sparsity coefficient chosen without ground truth

**Design.** Everywhere else our sparsity coefficient $\lambda$ is calibrated on the true latents (Appendix E). This experiment removes that access. On the ten confirmation draws per setting, our model is trained on sequences $0$ to $79$ of each draw and $\lambda$ is chosen on sequences $80$ to $99$. This is the released implementation's own split (its synthetic data configuration sets training identifiers $[0, 80]$

and validation identifiers $[80, 100]$), so our training sequences are exactly those the released implementation trains on. The configuration and recipe are those of Appendix F.1. For every $\lambda$ in the grid $\{5, 10, 20, 40, 80, 160, 320\}$, fixed in advance, shared by both settings and bracketing both calibrated values ($40$ and $150$), three model seeds are trained. The selection score (7) is the structural-model negative log-likelihood of the held-out sequences $\mathcal{N}_{\mathrm{v}} = \{80, \ldots, 99\}$, whose encoder means are standardized with the training statistics used by the graph stage. It contains no sparsity term and no other term that depends on $\lambda$. For each draw and seed the $\lambda$ with the lowest score is selected, and the per-draw value is the mean over the three seeds of the selected runs. The reconstruction term is left out because it sums over $10^4$ sites and barely changes with $\lambda$; the structural score is comparable across $\lambda$ only if the learned representation barely changes with $\lambda$, which is checked below.

**Results.** All 420 runs of our model and all 420 of the kernel finished. Our model outperforms the released implementation on all ten draws on F1 and on AUPRC in each setting, both at its default configuration and with its calibrated coefficient (Table 19). The selected $\lambda$ is spread over the grid (over the thirty pairs of draw and seed, the values $5, 10, \ldots, 320$ are chosen 3, 2, 9, 2, 8, 5, 1 times at $D = 10$ and 1, 8, 3, 4, 8, 5, 1 times at $D = 20$), so the score does not collapse to the densest graph. Within each draw and seed the latent MCC varies across $\lambda$ by at most $0.0094$ and the held-out reconstruction error by at most $1.5 \times 10^{-4}$, so the representation is nearly $\lambda$-invariant, as the score requires. Under the same selection the same-recipe coordinate kernel reaches mean F1 $0.600$ at $D = 10$, where our model wins all ten draws on F1 (mean gain $+0.069$); at $D = 20$ our model wins nine of ten draws on AUPRC ($p = 0.021$, mean gain $+0.148$).

Table 19: Sparsity coefficient chosen without ground truth, with the draw as the unit. Won, mean differences and sign-test $p$ are against the released implementation's default configuration. Won counts draws out of ten, as F1 / AUPRC. The last two columns count draws won against the released implementation with its calibrated coefficient and against the same-recipe coordinate kernel under the same selection.

| Setting | Valid | Won | Mean difference | Sign $p$ | Won vs calibrated | Won vs kernel |
|---|---|---|---|---|---|---|
| nonlinear, $D = 10$ | 10 | 10 / 10 | $+0.310$ / $+0.325$ | 0.002 / 0.002 | 10 / 10 | 10 / 8 |
| nonlinear, $D = 20$ | 10 | 10 / 10 | $+0.273$ / $+0.382$ | 0.002 / 0.002 | 10 / 10 | 7 / 9 |

Table 20: Sparsity coefficient chosen without ground truth, every draw. $\hat{\lambda}$ lists the selected coefficient of model seeds 0, 1, 2, and F1 and AUPRC are the means of the selected runs. Nearest is our F1 at the grid value nearest the calibrated coefficient ($40$ and $160$) on the same eighty sequences. Kernel is the F1 of the same-recipe coordinate kernel under the same selection, and the last two columns are the F1 of the released implementation at its default and calibrated coefficients.

| Draw | $\hat{\lambda}$ | F1 | AUPRC | Nearest | Kernel | Released | Calibrated |
|---|---|---|---|---|---|---|---|
| NL10 7011 | 320, 160, 80 | 0.737 | 0.732 | 0.727 | 0.663 | 0.270 | 0.258 |
| NL10 7012 | 5, 80, 20 | 0.674 | 0.688 | 0.702 | 0.593 | 0.293 | 0.316 |
| NL10 7013 | 20, 20, 20 | 0.830 | 0.844 | 0.793 | 0.827 | 0.490 | 0.419 |
| NL10 7014 | 80, 160, 10 | 0.784 | 0.791 | 0.786 | 0.719 | 0.412 | 0.357 |
| NL10 7015 | 80, 40, 20 | 0.615 | 0.631 | 0.592 | 0.589 | 0.148 | 0.154 |
| NL10 7016 | 20, 80, 5 | 0.610 | 0.710 | 0.625 | 0.560 | 0.356 | 0.316 |
| NL10 7017 | 40, 80, 20 | 0.597 | 0.615 | 0.592 | 0.509 | 0.537 | 0.581 |
| NL10 7018 | 20, 160, 160 | 0.777 | 0.741 | 0.740 | 0.702 | 0.343 | 0.323 |
| NL10 7019 | 80, 160, 80 | 0.486 | 0.367 | 0.453 | 0.264 | 0.200 | 0.222 |
| NL10 7020 | 5, 10, 20 | 0.580 | 0.691 | 0.679 | 0.569 | 0.542 | 0.242 |
| NL20 7021 | 160, 160, 10 | 0.720 | 0.717 | 0.739 | 0.574 | 0.302 | 0.294 |
| NL20 7022 | 10, 80, 5 | 0.544 | 0.603 | 0.585 | 0.676 | 0.292 | 0.359 |
| NL20 7023 | 10, 80, 40 | 0.498 | 0.652 | 0.519 | 0.517 | 0.340 | 0.342 |
| NL20 7024 | 10, 10, 10 | 0.529 | 0.651 | 0.595 | 0.606 | 0.283 | 0.269 |
| NL20 7025 | 20, 10, 160 | 0.579 | 0.675 | 0.644 | 0.566 | 0.435 | 0.424 |
| NL20 7026 | 160, 320, 160 | 0.590 | 0.534 | 0.590 | 0.508 | 0.405 | 0.370 |
| NL20 7027 | 80, 40, 80 | 0.643 | 0.673 | 0.606 | 0.547 | 0.386 | 0.346 |
| NL20 7028 | 80, 20, 10 | 0.484 | 0.644 | 0.555 | 0.424 | 0.271 | 0.197 |
| NL20 7029 | 40, 80, 20 | 0.643 | 0.695 | 0.645 | 0.596 | 0.276 | 0.165 |
| NL20 7030 | 80, 40, 80 | 0.762 | 0.759 | 0.720 | 0.591 | 0.272 | 0.295 |

**Ten further draws.** The same selection was applied to both factors on the ten further draws per setting of Appendix F.11, with the same grid, split and seeds. All runs finished. Table 21 gives the draws won by our factor against the coordinate kernel under this selection on the first ten, the further ten and all twenty draws, and Table 22 every further draw. Over all twenty draws the mean differences to the kernel are $+0.049$ F1 and $+0.085$ AUPRC at $D = 10$ and $+0.066$ F1 and $+0.166$ AUPRC at $D = 20$. Over the twenty draws of each setting our factor wins 17 in F1.

Table 21: Sparsity coefficient chosen without ground truth for both factors. Entries are the draws won by our factor against the same-recipe coordinate kernel (F1 / AUPRC) on the draws of Appendix F.1, on those of Appendix F.11 and over all twenty, with the sign-test $p$ for the twenty.

| Setting | First ten | Further ten | Twenty | Sign $p$ (twenty) |
|---|---|---|---|---|
| $D = 10$ | 10 / 8 | 7 / 8 | 17 / 16 | 0.003 / 0.012 |
| $D = 20$ | 7 / 9 | 10 / 10 | 17 / 19 | 0.003 / $4.0 \times 10^{-5}$ |

Table 22: Sparsity coefficient chosen without ground truth, the further draws: the three selected $\lambda$ (one per model seed) and the mean F1 / AUPRC over the selected runs, for our factor and the same-recipe coordinate kernel.

| Draw | Ours $\hat{\lambda}$ | Ours F1 / AUPRC | Kernel $\hat{\lambda}$ | Kernel F1 / AUPRC |
|---|---|---|---|---|
| NL10 7101 | 320, 320, 160 | 0.672 / 0.652 | 80, 160, 320 | 0.656 / 0.661 |
| NL10 7102 | 320, 160, 20 | 0.767 / 0.798 | 40, 320, 320 | 0.756 / 0.715 |
| NL10 7103 | 320, 160, 160 | 0.652 / 0.585 | 160, 10, 40 | 0.603 / 0.582 |
| NL10 7104 | 5, 10, 20 | 0.488 / 0.604 | 80, 80, 40 | 0.385 / 0.331 |
| NL10 7105 | 40, 5, 10 | 0.643 / 0.713 | 5, 320, 40 | 0.585 / 0.556 |
| NL10 7106 | 320, 80, 320 | 0.588 / 0.602 | 40, 20, 320 | 0.556 / 0.522 |
| NL10 7107 | 80, 40, 10 | 0.688 / 0.793 | 10, 10, 10 | 0.711 / 0.781 |
| NL10 7108 | 320, 320, 160 | 0.680 / 0.648 | 20, 320, 320 | 0.576 / 0.577 |
| NL10 7109 | 80, 40, 40 | 0.664 / 0.657 | 80, 80, 80 | 0.703 / 0.637 |
| NL10 7110 | 40, 80, 20 | 0.723 / 0.760 | 40, 160, 40 | 0.753 / 0.776 |
| NL20 7111 | 10, 40, 160 | 0.584 / 0.714 | 10, 80, 5 | 0.451 / 0.468 |
| NL20 7112 | 80, 160, 40 | 0.514 / 0.479 | 5, 10, 10 | 0.441 / 0.383 |
| NL20 7113 | 160, 80, 160 | 0.725 / 0.740 | 80, 80, 320 | 0.616 / 0.523 |
| NL20 7114 | 40, 160, 10 | 0.495 / 0.513 | 10, 40, 10 | 0.435 / 0.355 |
| NL20 7115 | 160, 80, 40 | 0.516 / 0.538 | 80, 40, 10 | 0.476 / 0.383 |
| NL20 7116 | 160, 160, 20 | 0.816 / 0.889 | 80, 80, 320 | 0.780 / 0.802 |
| NL20 7117 | 5, 320, 5 | 0.567 / 0.675 | 10, 320, 5 | 0.443 / 0.340 |
| NL20 7118 | 10, 40, 5 | 0.586 / 0.691 | 5, 20, 10 | 0.493 / 0.604 |
| NL20 7119 | 40, 40, 40 | 0.739 / 0.729 | 160, 160, 20 | 0.588 / 0.510 |
| NL20 7120 | 160, 20, 160 | 0.618 / 0.697 | 320, 10, 320 | 0.507 / 0.472 |

### F.10 Coordinate kernel with its own recipe selection

**Design.** Our recipe's prior weight $\beta = 0.05$ and its 10000 representation steps were selected with our factor. Its linear warm start over the first half of the graph schedule was set with the graph learner on true latents. All three settings were fixed before any fresh draw was generated. This experiment lets the same-recipe coordinate kernel select them for itself. On the calibration draws of Appendix F.6 (7001 at $D = 10$, 7042 at $D = 20$), never used in the confirmation, the kernel is trained over $\beta \in \{0.025, 0.05, 0.1, D/4\}$, 1200 or 10000 representation steps, and a linear warm start over 0 or half the schedule. It uses three model seeds, and all other settings and $\lambda$ are those of Appendix F.1. Per setting the configuration with the highest mean F1 is selected, with ties going to our recipe; this is the same kind of ground-truth access our recipe and the released implementation's calibrated coefficient had. If a setting selects our recipe, its confirmation values stand; otherwise the kernel is rerun with its selection on the ten confirmation draws.

**Results.** All 96 calibration and 30 confirmation runs finished. At $D = 10$ the kernel selects our recipe's settings, so the comparison of Appendix F.1 stands (nine of ten draws on F1). At $D = 20$ it selects $\beta = 0.1$ (Table 23); against the kernel with this selection our factor wins nine of ten draws on F1 (mean gain $+0.068$, sign test $p = 0.021$, Wilcoxon $p = 0.004$) and ten of ten on AUPRC ($+0.167$, $p = 0.002$) (Table 24). The kernel's mean F1 over the confirmation draws is $0.583$ with its

own selection and $0.591$ with our recipe, so the selection made on one draw does not transfer. Two components of the recipe help the kernel. Without the linear warm start the kernel is worse than with it at every other setting of the grid, on both calibration draws. At $\beta = D/4$ its representation collapses at $D = 20$ (MCC $0.71$ to $0.72$). The draw 7023, a near tie against the kernel with our recipe, is won against the kernel with its own selection ($0.563$ against $0.547$).

Table 23: Coordinate kernel with its own recipe selection: calibration grid, F1 / MCC means over three model seeds on the calibration draws. Bold marks the selected configuration; at $D = 10$ it is our recipe.

| $\beta$ | Representation steps | Linear warm start | $D = 10$ (draw 7001) | $D = 20$ (draw 7042) |
|---|---|---|---|---|
| 0.025 | 1200 | 0.0 | 0.622 / 0.995 | 0.401 / 0.908 |
| 0.025 | 1200 | 0.5 | 0.741 / 0.995 | 0.566 / 0.907 |
| 0.025 | 10000 | 0.0 | 0.577 / 0.995 | 0.439 / 0.907 |
| 0.025 | 10000 | 0.5 | 0.738 / 0.995 | 0.539 / 0.907 |
| 0.05 | 1200 | 0.0 | 0.595 / 0.995 | 0.461 / 0.907 |
| 0.05 | 1200 | 0.5 | 0.730 / 0.996 | 0.565 / 0.908 |
| 0.05 | 10000 | 0.0 | 0.564 / 0.995 | 0.477 / 0.907 |
| 0.05 | 10000 | 0.5 | **0.785 / 0.996** | 0.541 / 0.908 |
| 0.1 | 1200 | 0.0 | 0.626 / 0.995 | 0.447 / 0.907 |
| 0.1 | 1200 | 0.5 | 0.715 / 0.995 | 0.577 / 0.909 |
| 0.1 | 10000 | 0.0 | 0.562 / 0.996 | 0.511 / 0.910 |
| 0.1 | 10000 | 0.5 | 0.747 / 0.996 | **0.593 / 0.909** |
| $D/4$ | 1200 | 0.0 | 0.592 / 0.993 | 0.292 / 0.714 |
| $D/4$ | 1200 | 0.5 | 0.705 / 0.993 | 0.327 / 0.717 |
| $D/4$ | 10000 | 0.0 | 0.547 / 0.993 | 0.215 / 0.710 |
| $D/4$ | 10000 | 0.5 | 0.733 / 0.994 | 0.281 / 0.710 |

Table 24: Coordinate kernel with its own recipe selection, confirmation at $D = 20$: means over three model seeds. Ours and the kernel with our recipe are the values of Appendix F.1; at draw 7023 their F1 values are $0.5632$ and $0.5627$ before rounding.

| Draw | F1 ours | F1 kernel, our recipe | F1 kernel, own selection | AUPRC ours | AUPRC kernel, own selection |
|---|---|---|---|---|---|
| NL20 7021 | 0.774 | 0.609 | 0.615 | 0.745 | 0.554 |
| NL20 7022 | 0.659 | 0.681 | 0.674 | 0.614 | 0.586 |
| NL20 7023 | 0.563 | 0.563 | 0.547 | 0.639 | 0.476 |
| NL20 7024 | 0.676 | 0.658 | 0.652 | 0.686 | 0.609 |
| NL20 7025 | 0.647 | 0.625 | 0.616 | 0.691 | 0.476 |
| NL20 7026 | 0.606 | 0.543 | 0.526 | 0.564 | 0.409 |
| NL20 7027 | 0.602 | 0.568 | 0.562 | 0.625 | 0.463 |
| NL20 7028 | 0.547 | 0.477 | 0.449 | 0.600 | 0.324 |
| NL20 7029 | 0.714 | 0.572 | 0.554 | 0.671 | 0.445 |
| NL20 7030 | 0.723 | 0.617 | 0.632 | 0.696 | 0.521 |

### F.11 Ten further fresh draws per nonlinear setting

**Design.** Ten further draws per nonlinear setting (data seeds 7101 to 7110 at $D = 10$ and 7111 to 7120 at $D = 20$, not used before) are generated by the released generator exactly as in Appendix F.1. Our factor, the same-recipe coordinate kernel and the two-stage pipeline are trained on each with the settings of that appendix and three model seeds. The released implementation is not rerun. We compare ours with the same-recipe kernel on F1 and AUPRC, with the draw as the unit and a tie counted as not won. The comparison covers the ten new draws and the twenty draws pooled with Appendix F.1.

**Results.** All 180 runs finished. At $D = 20$ our model wins all ten new draws on F1 and on AUPRC, and over the twenty draws it wins 19 on F1 ($p = 4.0 \times 10^{-5}$) and 20 on AUPRC ($p = 1.9 \times 10^{-6}$). At $D = 10$ it wins nine of the ten new draws on AUPRC; over the twenty draws it wins 17 on AUPRC ($p = 0.003$), and its mean F1 gain is significant under the Wilcoxon signed-rank test ($p = 0.036$) (Table 25). Against the two-stage pipeline over the twenty draws, our model wins 14 draws on F1 and 16 on AUPRC at $D = 10$ (sign test $p = 0.115$ and $0.012$) and all twenty on both at $D = 20$ ($p = 1.9 \times 10^{-6}$). The pipeline's mean F1 and AUPRC over these draws are $0.608$ and $0.637$ at $D = 10$ and $0.486$ and $0.469$ at $D = 20$.

Table 25: Ten further fresh draws per nonlinear setting: ours minus the same-recipe coordinate kernel, with the draw as the unit. Entries are F1 / AUPRC. The last two columns pool these ten draws with the ten of Appendix F.1.

| Setting | Won (new ten) | Mean difference | Sign $p$ | Won (twenty) | Sign $p$ (twenty) |
|---|---|---|---|---|---|
| nonlinear, $D = 10$ | 5 / 9 | $+0.014$ / $+0.073$ | 1.000 / 0.021 | 14 / 17 | 0.115 / 0.003 |
| nonlinear, $D = 20$ | 10 / 10 | $+0.107$ / $+0.209$ | 0.002 / 0.002 | 19 / 20 | $4.0 \times 10^{-5}$ / $1.9 \times 10^{-6}$ |

Table 26: Ten further fresh draws per nonlinear setting, every draw (means over three model seeds). Ours is ISF, without smoothing at $D = 10$ and with smoothing at $D = 20$.

| Draw | F1 ours | F1 RBF | F1 PCA-VARIMAX | AUPRC ours | AUPRC RBF | MCC ours | MCC RBF |
|---|---|---|---|---|---|---|---|
| NL10 7101 | 0.662 | 0.621 | 0.667 | 0.728 | 0.725 | 0.987 | 0.969 |
| NL10 7102 | 0.790 | 0.803 | 0.731 | 0.876 | 0.843 | 0.974 | 0.977 |
| NL10 7103 | 0.565 | 0.579 | 0.579 | 0.608 | 0.505 | 0.954 | 0.916 |
| NL10 7104 | 0.548 | 0.394 | 0.318 | 0.588 | 0.372 | 0.976 | 0.853 |
| NL10 7105 | 0.663 | 0.662 | 0.708 | 0.720 | 0.560 | 0.976 | 0.900 |
| NL10 7106 | 0.577 | 0.569 | 0.638 | 0.599 | 0.559 | 0.986 | 0.899 |
| NL10 7107 | 0.716 | 0.746 | 0.614 | 0.811 | 0.755 | 0.989 | 0.984 |
| NL10 7108 | 0.706 | 0.649 | 0.598 | 0.782 | 0.706 | 0.980 | 0.980 |
| NL10 7109 | 0.671 | 0.698 | 0.617 | 0.679 | 0.619 | 0.994 | 0.981 |
| NL10 7110 | 0.740 | 0.779 | 0.645 | 0.787 | 0.804 | 0.996 | 0.996 |
| NL20 7111 | 0.665 | 0.533 | 0.479 | 0.721 | 0.411 | 0.991 | 0.897 |
| NL20 7112 | 0.473 | 0.451 | 0.427 | 0.468 | 0.303 | 0.928 | 0.864 |
| NL20 7113 | 0.748 | 0.662 | 0.595 | 0.775 | 0.567 | 0.996 | 0.935 |
| NL20 7114 | 0.522 | 0.446 | 0.418 | 0.464 | 0.276 | 0.941 | 0.828 |
| NL20 7115 | 0.544 | 0.476 | 0.461 | 0.543 | 0.341 | 0.963 | 0.881 |
| NL20 7116 | 0.845 | 0.779 | 0.536 | 0.868 | 0.788 | 0.997 | 0.980 |
| NL20 7117 | 0.655 | 0.486 | 0.484 | 0.675 | 0.350 | 0.989 | 0.833 |
| NL20 7118 | 0.740 | 0.642 | 0.563 | 0.754 | 0.628 | 0.996 | 0.958 |
| NL20 7119 | 0.751 | 0.542 | 0.541 | 0.704 | 0.384 | 0.998 | 0.906 |
| NL20 7120 | 0.692 | 0.542 | 0.515 | 0.747 | 0.587 | 0.946 | 0.893 |

### F.12 FREE NONNEGATIVE FOOTPRINTS

**Design.** This experiment compares our factor with footprints that are free rather than assembled from assigned units. Let $\mathcal{N}$ divide every column of a nonnegative matrix by its maximum, and let $\Psi$ be an unconstrained parameter matrix. The free nonnegative footprint is $F = \mathcal{N}\big(\mathrm{softplus}(\Psi)\big)$ with $\Psi \in \mathbb{R}^{L \times D}$, one free weight per site and mode and no incidence (FREEFOOTPRINT). The free weights on grid edges are $F = \mathcal{N}\big(\bar{H}\,\mathrm{softplus}(\Psi)\big)$ with $\Psi \in \mathbb{R}^{M \times D}$, nonnegative weights on the same grid-edge dictionary $\bar{H}$ as our factor but without a unit-to-mode assignment (FREEEDGES). Neither has an assignment sampler or a categorical KL term. The comparison is therefore between categorical unit assignment with its prior, heat-smoothed at $D = 20$, and unsmoothed free nonnegative weights, not an isolated effect of discreteness. On the twenty draws per setting of Appendices F.1 and F.11, both factors are trained with our recipe and model seeds 0, 1 and 2, with $\lambda = 40$ at $D = 10$ and $\lambda = 150$ at $D = 20$. The values of our factor are those of the two appendices.

**Results.** All runs finished. At $D = 20$, our factor outperforms the free nonnegative footprint on all twenty draws on F1 and on AUPRC. It outperforms the free weights on grid edges on 19 of 20 draws on F1 and on all twenty on AUPRC (Tables 27 and 28). At $D = 10$, where our factor has no smoothing (Section 6.2), the free variants reach its mean F1 and exceed its mean AUPRC. Our factor has the higher F1 on 12 and 10 of 20 draws and the higher AUPRC on 8 and 7 of 20 draws against the free footprint and the free weights on grid edges (Table 27).

Table 27: Free nonnegative footprints, twenty draws per setting: mean F1 / AUPRC, draws won by our factor, mean paired difference (ours minus the row) and sign-test $p$, each as F1 / AUPRC. Our factor is used without smoothing at $D = 10$ and with smoothing at $D = 20$ (Section 6.2).

| Setting | Factor | Mean F1 / AUPRC | Won | Mean difference | Sign $p$ |
|---|---|---|---|---|---|
| $D = 10$ | FREEFOOTPRINT | 0.665 / 0.708 | 12 / 8 | $+0.001$ / $-0.009$ | 0.503 / 0.503 |
| | FREEEDGES | **0.683** / **0.721** | 10 / 7 | $-0.016$ / $-0.022$ | 1.000 / 0.263 |
| | ISF (ours) | 0.666 / 0.699 | – | – | – |
| $D = 20$ | FREEFOOTPRINT | 0.562 / 0.564 | 20 / 20 | $+0.096$ / $+0.098$ | $1.9 \times 10^{-6}$ / $1.9 \times 10^{-6}$ |
| | FREEEDGES | 0.572 / 0.586 | 19 / 20 | $+0.086$ / $+0.076$ | $4.0 \times 10^{-5}$ / $1.9 \times 10^{-6}$ |
| | ISF (ours) | **0.657** / **0.662** | – | – | – |

Table 28: Free nonnegative footprints at $D = 20$, every draw (means over three model seeds).

| | F1 | | | AUPRC | | |
|---|---|---|---|---|---|---|
| Draw | Ours | FREEFOOTPRINT | FREEEDGES | Ours | FREEFOOTPRINT | FREEEDGES |
| NL20 7021 | 0.774 | 0.650 | 0.684 | 0.745 | 0.648 | 0.657 |
| NL20 7022 | 0.659 | 0.617 | 0.600 | 0.614 | 0.569 | 0.564 |
| NL20 7023 | 0.563 | 0.491 | 0.490 | 0.639 | 0.517 | 0.536 |
| NL20 7024 | 0.676 | 0.539 | 0.515 | 0.686 | 0.577 | 0.541 |
| NL20 7025 | 0.647 | 0.576 | 0.571 | 0.691 | 0.603 | 0.645 |
| NL20 7026 | 0.606 | 0.552 | 0.542 | 0.564 | 0.542 | 0.537 |
| NL20 7027 | 0.602 | 0.519 | 0.516 | 0.625 | 0.452 | 0.506 |
| NL20 7028 | 0.547 | 0.493 | 0.487 | 0.600 | 0.579 | 0.565 |
| NL20 7029 | 0.714 | 0.476 | 0.501 | 0.671 | 0.508 | 0.573 |
| NL20 7030 | 0.723 | 0.633 | 0.604 | 0.696 | 0.615 | 0.616 |
| NL20 7111 | 0.665 | 0.589 | 0.649 | 0.721 | 0.669 | 0.704 |
| NL20 7112 | 0.473 | 0.457 | 0.481 | 0.468 | 0.389 | 0.426 |
| NL20 7113 | 0.748 | 0.634 | 0.619 | 0.775 | 0.652 | 0.654 |
| NL20 7114 | 0.522 | 0.450 | 0.486 | 0.464 | 0.325 | 0.415 |
| NL20 7115 | 0.544 | 0.480 | 0.525 | 0.543 | 0.484 | 0.489 |
| NL20 7116 | 0.845 | 0.700 | 0.726 | 0.868 | 0.741 | 0.745 |
| NL20 7117 | 0.655 | 0.521 | 0.546 | 0.675 | 0.515 | 0.565 |
| NL20 7118 | 0.740 | 0.652 | 0.673 | 0.754 | 0.657 | 0.685 |
| NL20 7119 | 0.751 | 0.659 | 0.674 | 0.704 | 0.633 | 0.670 |
| NL20 7120 | 0.692 | 0.544 | 0.544 | 0.747 | 0.608 | 0.628 |

### F.13 HEAT KERNEL AND SINGLE-PARENT DECODING

On the twenty fresh draws per setting of Appendices F.1 and F.11, the hypergraph heat kernel (HEATKERNEL) and single-parent decoding (SINGLEPARENT) are trained with our recipe, $\lambda^\star$ and model seeds 0, 1 and 2. All runs finished. Table 29 gives the comparisons of our model with HEATKERNEL and SINGLEPARENT.

Table 29: Heat kernel and single-parent decoding on the twenty fresh draws per setting: draws won by our model (F1 / AUPRC), mean paired difference (ours minus the row) and sign-test $p$ (F1).

| Setting | Comparator | Won | Mean difference | Sign $p$ |
|---|---|---|---|---|
| $D = 10$ | HEATKERNEL | 17 / 19 | $+0.069$ / $+0.082$ | 0.003 |
| $D = 10$ | SINGLEPARENT | 20 / 20 | $+0.174$ / $+0.210$ | $1.9 \times 10^{-6}$ |
| $D = 20$ | HEATKERNEL | 20 / 20 | $+0.171$ / $+0.242$ | $1.9 \times 10^{-6}$ |
| $D = 20$ | SINGLEPARENT | 20 / 20 | $+0.288$ / $+0.361$ | $1.9 \times 10^{-6}$ |

## G ABLATIONS

This appendix lists ablations of ISF. Each removes or replaces one or more parts of the model or of its input, while the data, the training configuration and the sparsity coefficient stay fixed. Table 30 names each ablation, the draws or structures on which ISF has the higher value, and where it is reported. ISF outperforms every variant listed except three, which match or exceed it: the two free variants at $D = 10$ on the grid benchmark (Appendix F.12) and continuous loadings on the hyperedge dictionary of the committee hypergraphs (Table 32). Table 31 gives the smoothing ablation in both grid settings.

Table 30: Ablations of ISF. The third column gives the draws or structures on which ISF has the higher value, of 20 draws per grid setting (10 for the rewiring) and 10 structures per committee or anchored setting. On the grid, ISF is the configuration of Table 1, which smooths at $D = 20$ only, except in the smoothing row, where it is the variant with smoothing. On the committee hypergraphs and anchored structures, ISF is unsmoothed.

| Removed or replaced | Variant | ISF higher on | Details |
|---|---|---|---|
| *Grid benchmark* | | | |
| Smoothing | ISF without smoothing | F1 16, 19; AUPRC 13, 20 ($D = 10, 20$) | Table 31 |
| Assignment, smoothing | free weights on grid edges | F1 19; AUPRC 20 ($D = 20$) | Table 2 |
| Incidence, assignment, smoothing | free nonnegative footprint | F1 20; AUPRC 20 ($D = 20$) | Table 2 |
| Assignment | free weights on grid edges | F1 10; AUPRC 7 ($D = 10$) | App. F.12 |
| Incidence, assignment | free nonnegative footprint | F1 12; AUPRC 8 ($D = 10$) | App. F.12 |
| Correct incidence | rewired grid adjacency | F1 10; AUPRC 10 (both $D$) | App. F.8 |
| Atoms | heat kernel, same recipe | F1 17, 20; AUPRC 19, 20 ($D = 10, 20$) | App. F.13 |
| Multi-parent assignment | single-parent decoding | F1 20; AUPRC 20 (both $D$) | App. F.13 |
| *Committee hypergraphs (C20, C15, C20N10)* | | | |
| Hyperedge atoms | pair atoms of the clique expansion | MCC 10 in each setting | App. F.2 |
| Correct memberships | rewired memberships | MCC 10 in each setting | App. F.2 |
| Atoms | heat kernel, same pairwise operator | MCC 10 in each setting | App. F.2 |
| Assignment | continuous loadings, same hyperedges | MCC 2 and 1 (C20, C15) | Table 32 |
| *Anchored structures* | | | |
| Learned observation maps | identity map | MCC 10 | App. F.5 |

**Unit assignment and continuous loadings.** This ablation keeps the hyperedge dictionary of the committee hypergraphs and replaces the unit-to-mode assignment by continuous loadings, $F = \mathcal{N}\big(\bar{H}\,\mathrm{softplus}(\Psi)\big)$ with $\Psi \in \mathbb{R}^{M \times D}$ and $\mathcal{N}$ as in Appendix F.12. It has no assignment sampler and no categorical KL term, and every other part of the model, the training configuration and $\lambda$ are those of Appendix F.2. Both arms were rerun together on the structures of settings C20 and C15 with model seeds 0, 1 and 2. The mean MCC of ISF in this run differs from Table 3 by at most 0.001, and its mean F1 and AUPRC by at most 0.02. The continuous loadings reach a higher latent MCC

Table 31: Smoothing ablation on the forty fresh draws of the grid benchmark, with the recipe and $\lambda$ of Table 1. Columns give the mean F1 / AUPRC of ISF without and with the smoothing (5), the draws won by the variant with smoothing and the sign-test $p$, each as F1 / AUPRC. The configuration of Table 1 is the variant without smoothing at $D = 10$ and the variant with smoothing at $D = 20$. It was fixed before these draws were generated, and the other variant was run afterwards on the same draws. The table uses no bold.

| Setting | Without smoothing | With smoothing | Won with smoothing | Sign $p$ |
|---|---|---|---|---|
| nonlinear, $D = 10$ | 0.666 / 0.699 | 0.714 / 0.732 | 16 / 13 | 0.012 / 0.263 |
| nonlinear, $D = 20$ | 0.551 / 0.534 | 0.657 / 0.662 | 19 / 20 | $4.0 \times 10^{-5}$ / $1.9 \times 10^{-6}$ |

than the unit assignment on 8 of 10 structures in C20 and on 9 of 10 in C15, and a higher mean edge F1 and AUPRC in both settings (Table 32).

Table 32: Unit assignment (ISF) against continuous nonnegative loadings on the same hyperedge dictionary, committee hypergraphs, ten structures per setting. Entries are means over three model seeds as MCC / F1 / AUPRC, the structures on which ISF has the higher value, and the sign-test $p$ for MCC. In C15 the two arms have exactly equal F1 on one structure, which counts as not won.

| Setting | ISF | Continuous loadings | ISF higher (MCC / F1 / AUPRC) | Sign $p$ (MCC) |
|---|---|---|---|---|
| C20 | 0.790 / 0.541 / 0.508 | **0.818 / 0.631 / 0.608** | 2 / 3 / 1 | 0.109 |
| C15 | 0.882 / 0.698 / 0.671 | **0.896 / 0.768 / 0.762** | 1 / 2 / 2 | 0.021 |

## H PARAMETER STUDIES

Table 33 varies the sparsity coefficient $\lambda$ over the grid of Appendix F.9 for ISF and the same-recipe coordinate kernel. ISF has the higher mean F1 and AUPRC at every value in both settings and wins most draws at every value. Further parameter studies appear with their experiments. Appendix E varies the sparsity coefficient of the graph learner on the true latents, and Appendix F.6 that of the released implementation. Appendix F.10 varies the representation settings of the coordinate kernel. The committee settings C20, C15 and C20N10 vary the observation noise and the number of sequences (Appendix F.2).

Table 33: Sensitivity to the sparsity coefficient under the protocol of Appendix F.9, with training sequences 0 to 79, twenty draws per setting and three model seeds. Columns give the mean F1 / AUPRC of ISF and of the same-recipe coordinate kernel, and the draws won by ISF (F1 / AUPRC). These runs train on fewer sequences than those of Table 1, so their values differ from it. The table uses no bold. The mean latent MCC changes by at most $0.001$ across $\lambda$ (ISF $0.975$ at $D = 10$ and $0.971$ at $D = 20$; kernel $0.938$ and $0.904$).

| | nonlinear, $D = 10$ | | | nonlinear, $D = 20$ | | |
|---|---|---|---|---|---|---|
| $\lambda$ | ISF | Kernel | Won | ISF | Kernel | Won |
| 5 | 0.599 / 0.712 | 0.583 / 0.632 | 13 / 15 | 0.502 / 0.672 | 0.467 / 0.507 | 16 / 18 |
| 10 | 0.631 / 0.711 | 0.605 / 0.636 | 12 / 15 | 0.543 / 0.690 | 0.504 / 0.521 | 16 / 20 |
| 20 | 0.657 / 0.712 | 0.622 / 0.629 | 15 / 18 | 0.585 / 0.694 | 0.537 / 0.524 | 16 / 18 |
| 40 | 0.670 / 0.697 | 0.628 / 0.623 | 14 / 15 | 0.626 / 0.688 | 0.566 / 0.512 | 17 / 18 |
| 80 | 0.668 / 0.672 | 0.630 / 0.594 | 12 / 14 | 0.639 / 0.665 | 0.571 / 0.493 | 16 / 18 |
| 160 | 0.674 / 0.650 | 0.610 / 0.565 | 17 / 17 | 0.635 / 0.626 | 0.566 / 0.464 | 16 / 18 |
| 320 | 0.612 / 0.581 | 0.556 / 0.504 | 15 / 15 | 0.628 / 0.585 | 0.540 / 0.418 | 18 / 18 |

## I IMPLICATIONS FOR LEARNING ON HYPERGRAPHS

Two consequences of the results reach beyond latent causal discovery. First, any model whose only structural input is $HW\mathrm{diag}(\delta)^{-1}H^\top$, such as heat kernels and spectral convolutions on the

normalized hypergraph Laplacian (Zhou et al., 2006; Feng et al., 2019), gives identical outputs on incidences with the same weighted clique expansion and degrees. By Theorem 2, such incidences can carry different latent coordinates, so within the model class of Theorem 2 the pairwise operator is not sufficient. Second, committee hypergraphs with membership-following footprints provide controlled comparisons of unit-level and pair-level parameterizations, and their anchored variants a setting in which (C1) and (C2) can be verified.

## J RUNTIME ACCOUNTING

All experiments ran on NVIDIA GeForce RTX 5090 cards (32 GB) of already-running hosts: the calibration of Appendix E on one host and the experiments of Appendix F on three further hosts. In the runs of Appendix F.11, three runs of our model or of a same-recipe comparator share one card. A run of our model takes a median of $340$ seconds at $D = 10$ and $360$ seconds at $D = 20$, and a run of the same-recipe coordinate kernel $305$ seconds. A 400-epoch run of the released implementation, alone on one card, takes a median of $82$ minutes. Its $20$ runs in Appendix F.1, whose start and end were logged on one host, range from $81$ to $96$ minutes.