# LEARNING HIGHER-ORDER INTERACTIONS FROM INCOMPLETE SPATIOTEMPORAL DATA WITH MIXTURES OF HYPERGRAPH LAPLACIANS

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

## ABSTRACT

Traffic sensor data are strongly structured in time and across sensors, but reconstructing missing values becomes increasingly difficult as gaps lengthen because missingness removes both local temporal context and the joint observations needed to infer cross-sensor relationships. We introduce **MoHL**, a mixture of hypergraph Laplacians that learns group-wise relational structure directly from incomplete spatiotemporal data. MoHL discovers overlapping sensor groups from observed co-variation, assigns each interaction order a Laplacian component, and learns their mixture from mask-matched holdouts. This group-wise parameterization pools evidence across related sensors rather than estimating every coupling independently. We show that the learned mixture induces structured weight sharing and characterize when such sharing is beneficial. Across three traffic networks, higher-order groups provide the largest gains under extended gaps, and shared weights consistently outperform freely learned weights on the same support when relational evidence is scarce. MoHL achieves the lowest error among evaluated methods under day-long gaps while using orders of magnitude fewer learned parameters than deep imputers. These results show that higher-order group structure is most useful precisely when missingness makes relational evidence scarce.

## 1 INTRODUCTION

Traffic networks are widely instrumented with roadside sensors, such as inductive-loop and radar detectors, that continuously measure traffic speed, flow, and occupancy (Klein et al., 2006). These measurements support traffic monitoring, forecasting, and control (Li et al., 2018), but are often incomplete because of equipment faults, communication failures, and other disruptions (Fekade et al., 2017; Chen et al., 2018). Recovering missing measurements is therefore essential for maintaining a reliable view of network conditions. Traffic sensor data are strongly structured in both time and space: individual sensors follow recurring temporal patterns, while sensors exposed to common demand, congestion, or traffic propagation often vary together (Yin et al., 2015; Li et al., 2020). Such dependence can also occur at the group level, with multiple sensors exhibiting coordinated deviations from their typical behavior. These relationships become especially valuable during long outages, yet the same outages reduce the joint observations available to learn them. The central challenge is therefore to recover reliable relational structure from incomplete evidence.

How should an imputer capture and represent this structure? A natural first choice is the road graph, where edges encode physical connectivity or proximity. Yet road topology alone may miss dependencies induced by shared demand and evolving traffic conditions, motivating methods that learn spatial relationships directly from data (Xu et al., 2024; Bai et al., 2020; Wang et al., 2025). Methods that rely on fixed graphs may therefore inherit an incomplete relational prior. Learning pairwise relationships directly is more flexible, but estimating a separate weight for every sensor pair requires sufficient joint observations for hundreds or thousands of couplings, precisely the evidence that structured missingness removes. This motivates learning relationships at the group level, where related couplings can share statistical evidence.

We use hypergraphs to represent these group relationships. A hyperedge links sensors expected to vary together and ties multiple pairwise couplings through one group relation (Zhou et al., 2006;



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

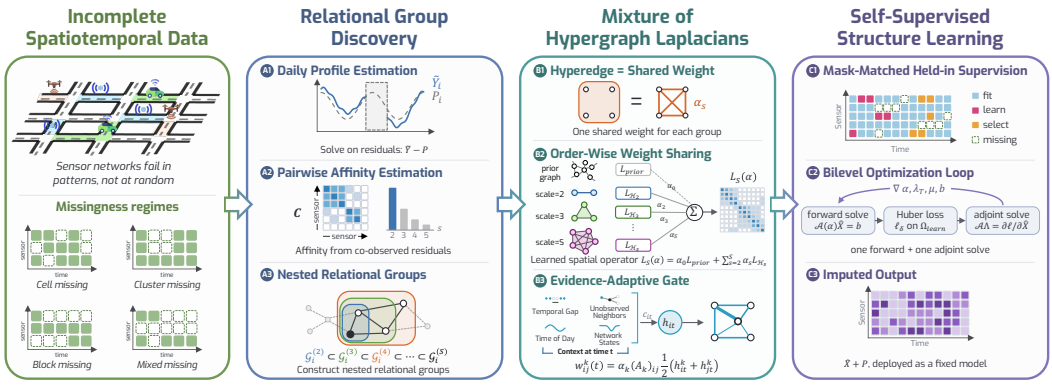


Figure 1: **Overview of MoHL**: affinities computed from incomplete observations define relational groups; a learned mixture of their hypergraph Laplacians enters a spatiotemporal imputation solve.

Feng et al., 2019; Bai et al., 2021; Chien et al., 2022). For the quadratic Laplacian used here, this means that the induced graph edges share weights. Rather than estimating every pair independently, a group pools evidence across its member pairs. This can reduce estimation variance when relational evidence is scarce, while free pairwise weights become preferable when evidence is abundant or the groups are inaccurate.

We introduce **MoHL**, a *mixture of hypergraph Laplacians* that learns these group relationships from incomplete data (Figure 1). MoHL removes each sensor's typical daily pattern, estimates affinities from observed pairs of residuals, and builds overlapping groups at multiple interaction orders. Each order contributes one Laplacian component with a shared coefficient learned from mask-matched holdouts. Their mixture defines the spatial regularization in a positive-definite spatiotemporal imputation solve, with an evidence-adaptive gate adjusting the regularization to local missingness.

Our analysis asks when and why this shared parameterization helps. We characterize the pairwise couplings induced by the mixture and give a Gaussian group model under which sharing is Bayes-optimal. To isolate sharing itself, we compare against freely learned pairwise weights on the same support and vary the observation window to control the evidence available per coupling. Higher-order groups help most during long gaps, and shared weights outperform free ones when joint evidence is scarce. As evidence grows, the advantage narrows, and it depends on discovering meaningful groups.

**Contributions.** **(1) Group-wise structure from incomplete data.** MoHL discovers overlapping multiscale sensor groups directly from incomplete observations and represents them with a mixture of hypergraph Laplacians. **(2) Learnable mixture across interaction orders.** MoHL assigns one Laplacian component to each group order and learns how much each order should contribute from mask-matched holdouts. **(3) Evidence-aware validation of group sharing.** Theory and controlled experiments show that group-wise sharing is most effective when relational evidence is scarce.

## 2 Problem Statement and Preliminaries

### 2.1 Problem Statement

We observe $N$ sensors at $T$ regular times; $X^* \in \mathbb{R}^{N\times T}$ holds their true values, and each entry $(i,t)$ is a *cell*. The mask $M \in \{0,1\}^{N\times T}$ marks cells as observed ($M_{i,t}=1$) or missing ($M_{i,t}=0$), with $\Omega := \{(i,t) : M_{i,t}=1\}$ the observed cells and $\Omega^c$ the missing ones. An observed cell carries a noisy reading $Y^{\text{obs}}_{i,t} = X^*_{i,t} + \xi_{i,t}$, with $\xi_{i,t} \sim \text{subG}(\sigma^2)$ and observation propensity $\pi_{i,t} := \mathbb{E}[M_{i,t}] \geq \pi_{\min} > 0$. Missing cells need not be scattered: a gap can span hours or days of one sensor, or several nearby sensors at once. Given $Y = M \odot Y^{\text{obs}}$, a prior sensor adjacency $A$, inter-sensor distances and the sampling times, the task is to impute $X^*$ on $\Omega^c$ from the observed cells alone, scored by mean absolute error against withheld truth.

**Missingness removes temporal and relational evidence.** Two quantities separate an isolated missing cell from a long gap. The *temporal gap* $g_{i,t} := \min\{|t - t'| : M_{i,t'} = 1\}$ is the distance to sensor $i$'s nearest observation, infinite if the sensor is never observed; the *relational evidence* $n_{ij} := \sum_t M_{i,t} M_{j,t}$ counts the times sensors $i$ and $j$ are observed together. Short gaps leave local temporal context; long gaps remove it, forcing the estimate onto other sensors while shrinking the joint observations from which those relations are learned. The regimes that need relational structure most thus provide the least evidence for it.

### 2.2 Hypergraphs and hypergraph Laplacians

**From pairs to groups.** A weighted graph represents sensor relationships pairwise: with adjacency $A \geq 0$ and degree matrix $D_A := \mathrm{diag}(A\mathbf{1})$, the Laplacian $D_A - A$ penalizes differences between connected sensors. This pairwise view has two limitations. A fixed graph such as road proximity may miss sensors that respond to the same traffic conditions, while learning all pairwise weights among $s$ related sensors requires up to $\binom{s}{2}$ parameters. A hypergraph $\mathcal{H} = (V, E)$ instead represents the group directly. Its nodes are the sensors $V = [N]$, and each hyperedge $e \in E$ joins $|e| \geq 2$ sensors with one weight $w_e \geq 0$. We call $|e|$ the interaction *order*; an ordinary graph is the order-two case.

**Hypergraph Laplacian.** With $X_{i,:}$ the time series of sensor $i$ and $\bar{X}_{e,:} := |e|^{-1} \sum_{i \in e} X_{i,:}$ the group average, a hyperedge penalizes disagreement with that average:

$$\langle X, P_e X \rangle_F \;=\; \sum_{i \in e} \left\| X_{i,:} - \bar{X}_{e,:} \right\|_2^2, \qquad P_e \;:=\; I_e \;-\; \tfrac{1}{|e|}\, \mathbf{1}_e \mathbf{1}_e^\top, \qquad L_{\mathcal{H}} \;:=\; \sum_{e \in E} w_e\, P_e. \quad (1)$$

Here $I_e$ is the identity and $\mathbf{1}_e$ the indicator on $e$. The projector $P_e$ leaves motion common to the group unpenalized and has rank $|e| - 1$; for $e = \{i, j\}$ it is half the usual pairwise Laplacian.

**Hypergraph Laplacians share evidence within groups.** Group disagreement is also the average of pairwise disagreements, $\sum_{i \in e} \|X_{i,:} - \bar{X}_{e,:}\|_2^2 = |e|^{-1} \sum_{i<j;\, i,j \in e} \|X_{i,:} - X_{j,:}\|_2^2$, so $P_e = D_{A^{(e)}} - A^{(e)}$ is the Laplacian of the clique $A^{(e)}$ with weight $1/|e|$ on every pair of $e$. By linearity of the degree matrix, $L_{\mathcal{H}} = \sum_e w_e \big(D_{A^{(e)}} - A^{(e)}\big) = D_{A_{\mathcal{H}}} - A_{\mathcal{H}}$ with induced adjacency $(A_{\mathcal{H}})_{ij} = \sum_{e \ni i,j} w_e/|e|$ (Lemma 3). A hyperedge therefore ties the couplings of all sensor pairs in the group through one shared weight, pooling their evidence rather than estimating each relation independently. This is the statistical role of higher-order structure in MoHL.

**MoHL mixes interaction orders.** One weight per discovered group could still be costly. As a design choice that simplifies learning, MoHL builds overlapping groups around each sensor at increasing orders (Section 3.1), shares one weight across all groups of an order, and keeps the prior graph as a component. With $\mathcal{H}_s$ the order-$s$ groups, $L_{\mathcal{H}_s} := \rho_s^{-1} \sum_{e \in \mathcal{H}_s} P_e$ and $L_{\mathrm{prior}} := \rho_0^{-1}(D_A - A)$, scaled by the largest eigenvalues $\rho_s$ of $\sum_{e \in \mathcal{H}_s} P_e$ and $\rho_0$ of $D_A - A$, the spatial operator is

$$L_S(\alpha) \;=\; \alpha_0\, L_{\mathrm{prior}} \;+\; \sum_{s=2}^{S} \alpha_s\, L_{\mathcal{H}_s}, \qquad \alpha \geq 0. \quad (2)$$

The mixture uses one coefficient for the prior graph and one for each interaction order, keeping the number of learned structural weights small while allowing different group scales to contribute differently. Setting $\alpha_s = 0$ for all $s \geq 2$ recovers the prior graph. Because the mixture can downweight an unreliable group structure but cannot change its membership, the next step is to discover groups reliably from incomplete observations.

## 3 MoHL: Learning Mixtures of Hypergraph Laplacians

MoHL separates two questions: *which sensors belong together*, and *how much should each interaction order matter*? We first discover multiscale groups from the observations that remain, then learn the mixture of their Laplacians from held-out data. An evidence-adaptive gate further adjusts the learned structure to local missingness.

### 3.1 RELATIONAL INTERACTION GROUPS FROM INCOMPLETE DATA

**Remove recurring temporal patterns.** Sensors may share a daily cycle without responding similarly to changing traffic conditions. We therefore discover relationships from residual behavior. Let $\tilde{Y}$ denote per-sensor standardized observations. For sensor $i$ and 30-minute time-of-day bin $b$, $P_i(b) := \mathrm{mean}\{\tilde{Y}_{i,t} : M_{i,t} = 1,\ b(t) = b\}$. After smoothing and interpolation, $P_{i,t}$ defines the daily profile. Groups are discovered from residuals $\tilde{Y} - P$, and $P$ is added back after imputation. Profiles are always computed from the observations available to the current solve.

**Estimate affinity from observed pairs.** Requiring all members of a group to be observed together would waste much of the remaining data. Instead, let $\bar{R}$ be the masked one-hour moving average of $(\tilde{Y} - P) \odot M$, and define

$$C_{ij} := \mathrm{corr}\big(\bar{R}_{i,\cdot},\, \bar{R}_{j,\cdot}\big)\big|_{\{t:\ \bar{R}_{it},\, \bar{R}_{jt} \text{ defined}\}}, \qquad C_{ij} := 0 \text{ if fewer than } n_{\min} \text{ timesteps are shared} \tag{3}$$

At observation rate $\pi$, a pair has $T\pi^2$ expected joint observations, compared with $T\pi^s$ for an order-$s$ group (Lemma 5). Pairwise affinities therefore provide a more reliable basis for discovering higher-order groups under missingness.

**Build nested groups.** For each observed sensor, we rank neighbors by affinity and add them progressively, producing overlapping groups across interaction orders.

**Definition 1** (Relational interaction groups)**.** For anchor $i$ and order $s \geq 2$, let $\mathcal{G}_i^{(s)} = \{i\} \cup \mathrm{Top}_{s-1}(C_{i\cdot})$, where $\mathrm{Top}_{s-1}$ selects the $s-1$ highest-affinity neighbors. Thus $\mathcal{G}_i^{(2)} \subset \cdots \subset \mathcal{G}_i^{(S)}$. A group is retained only if every selected neighbor has affinity at least $c_{\min}$ to the anchor. After removing duplicates, $\mathcal{H}_s$ denotes the retained order-$s$ groups.

The nesting grows each local relation from its strongest members, while the threshold excludes weak groups that would otherwise impose incorrect smoothing.

### 3.2 THE MOHL ESTIMATOR

**Spatiotemporal imputation.** Given a positive-semidefinite spatial operator $L_S$ and temporal difference Laplacian $L_T$, we recover the missing values by

$$\widehat{X}(L_S) \;:=\; \arg\min_X \;\; \mathcal{F}(X) \;+\; \lambda_S\, \mathcal{R}_S(X) \;+\; \lambda_T\, \mathcal{R}_T(X) \;+\; \tfrac{\mu}{2}\|X\|_F^2. \tag{4}$$

where $\mathcal{F}(X) = \frac{1}{2}\|M \odot (X - Y)\|_F^2$ fits observed cells, $\mathcal{R}_S(X) = \frac{1}{2}\langle X, L_S X\rangle_F$ enforces spatial consistency, and $\mathcal{R}_T(X) = \frac{1}{2}\langle X, X L_T\rangle_F$ enforces temporal smoothness. The ridge $\mu > 0$ guarantees a unique solution. In implementation, the fidelity term is weighted by inverse observation propensity.

**Insert the learned groups.** Substitute the mixture (2) for $L_S$, absorb $\lambda_S$ into $\alpha$, and solve for residuals $\tilde{Y} - P$. With $W$ the propensity-weighted mask, the normal equations are

$$\mathcal{A}(\alpha, \lambda_T, \mu)\, \hat{X} \;=\; W \odot (\tilde{Y} - P), \qquad \mathcal{A}Z \;:=\; W \odot Z + L_S(\alpha)\, Z + \lambda_T\, Z L_T + \mu Z. \tag{5}$$

Because $\mathcal{A}$ is symmetric positive definite, the solution is unique and can be computed efficiently by conjugate gradients. All spatial structure enters through $L_S(\alpha)$, which lets us compare different group structures within the same estimator.

**What the mixture can express.** The $S$ structure weights $\alpha_s = e^{a_s}$ include one prior weight and one per group order; with $\lambda_T$ and $\mu$, this is eight scalars when $S = 6$. Each $\alpha_s$ contributes to every pair in every retained order-$s$ group. Overlapping groups add their contributions. The following result characterizes the induced pairwise operator.

**Theorem 1** (Structure and identifiability of the shared-weight operator)**.** *Let* $K_{ij}^{(s)} := |\{e \in \mathcal{H}_s : \{i, j\} \subseteq e\}|$ *and let* $\varphi(i, j) := (A_{ij}, K_{ij}^{(2)}, \ldots, K_{ij}^{(S)})$ *be the pair's* profile*. For every* $\alpha \geq 0$*: (i)*

$L_S(\alpha)$ *is the graph Laplacian of the weighted adjacency*

$$A(\alpha)_{ij} \;=\; \frac{\alpha_0}{\rho_0}\,A_{ij} \;+\; \sum_{s=2}^{S} \frac{\alpha_s}{s\,\rho_s}\,K^{(s)}_{ij} \qquad (i \neq j;\ \rho_0, \rho_s \textit{ as in } (2)); \tag{6}$$

(ii) *two sensor pairs receive the same coupling for every* $\alpha$ *if and only if their profiles* $\varphi(i,j)$ *are identical;* (iii) *if some prior edge belongs to no group and each nonempty order introduces a new non-prior pair, then* $\alpha \mapsto L_S(\alpha)$ *is injective.*

Thus, the discovered groups determine which sensor relations share parameters, while the mixture weights determine how strongly each interaction order contributes. Under the stated conditions, these order weights are identifiable.

**Why can sharing help?** A hyperedge also has a probabilistic interpretation: its members fluctuate around a shared latent group level.

**Theorem 2** (Latent-factor representation)**.** *(a) For a hyperedge* $e$ *and* $\alpha > 0$, $\exp(-\frac{\alpha}{2}x^\top P_e x) \propto \int_{\mathbb{R}} \exp\!\big(-\frac{\alpha}{2}\sum_{i\in e}(x_i - m)^2\big)\,dm$*: a hyperedge penalty is the marginal likelihood of* $x_i = m_e + \varepsilon_i$ *with* $\varepsilon_i \sim \mathcal{N}(0, 1/\alpha)$*, with a flat prior on* $m_e$*. (b) Under the corresponding Gaussian spatiotemporal model, with observation noise precision* $W$*, the solution of* (5) *at the true parameters is the posterior mean and median.*

The theorem explains the role of group sharing: sensors in a valid group borrow strength through a common latent behavior. This can reduce estimation variance when evidence is scarce, but incorrect or heterogeneous groups introduce bias. Section 3.4 tests this trade-off against free weights on the same support. Proofs are in Appendix D.

### 3.3 Self-supervised learning of the mixture

**Learn from observed holdouts.** Missing cells provide no labels, so MoHL learns from observed cells. We split $\Omega$ into fitting cells $\Omega_{\text{fit}}$ and held-out cells $\Omega_{\text{learn}}$, solve on $\Omega_{\text{fit}}$, and optimize prediction on $\Omega_{\text{learn}}$:

$$\min_{a,\,\log\lambda_T,\,\log\mu}\ \frac{1}{|\Omega_{\text{learn}}|}\sum_{(i,t)\in\Omega_{\text{learn}}} \ell_\delta\big(\hat{X}_{it} - (\tilde{Y}_{it} - P_{it})\big) + \gamma\sum_{s\geq 2} e^{a_s} \quad \text{s.t. (5) on } \Omega_{\text{fit}}, \tag{7}$$

where $\alpha_s = e^{a_s}$ ensures nonnegative mixture weights. The penalty suppresses higher orders with little predictive support. Profiles, groups, and the solve use only $\Omega_{\text{fit}}$.

**Differentiate through one solve.** The weights affect held-out error through (5). Instead of differentiating every conjugate-gradient iteration, we solve one adjoint system $\mathcal{A}\Psi = \partial\ell/\partial\hat{X}$. Since $\mathcal{A}$ is symmetric, this uses the same operator as the forward solve and gives every parameter gradient by an inner product (Proposition 3):

$$\frac{\partial\ell}{\partial a_s} = -\alpha_s\langle\Psi,\, L_{\mathcal{H}_s}\hat{X}\rangle_F, \qquad \frac{\partial\ell}{\partial\log\lambda_T} = -\lambda_T\langle\Psi,\,\hat{X}L_T\rangle_F, \qquad \frac{\partial\ell}{\partial\log\mu} = -\mu\langle\Psi,\,\hat{X}\rangle_F. \tag{8}$$

The regularizer adds $\gamma\alpha_s$ to the derivative for $s \geq 2$. Each learning step needs one forward and one adjoint solve regardless of the number of orders; Appendix D.7 gives the derivation.

**Match holdouts to deployment gaps.** Random holdouts mostly teach short-gap interpolation. We instead sample holdouts that mimic the observed missingness pattern, including long intervals and dark sensors. Alternating day blocks form $\Omega_{\text{learn}}$ and a separate selection set $\Omega_{\text{sel}}$. Profiles and groups are rebuilt after each holdout to prevent leakage.

**Adapt to local evidence.** The same mixture need not suit every missing cell. We therefore gate each component using observed context $c_{it}$, including temporal gap, missing neighbors, daily harmonics, and network-mean speed. For component $k$ with adjacency $A_k$ ($A/\rho_0$ for the prior graph, $K^{(s)}/(s\rho_s)$ for order $s$), the edge weight at time $t$ is

$$w^k_{ij}(t) = \alpha_k\,(A_k)_{ij}\,\tfrac{1}{2}\big(h^k_{it} + h^k_{jt}\big), \qquad h^k_{it} = \exp\big(\text{clip}(c_{it}^\top b_k, -\kappa, \kappa)\big), \tag{9}$$

The symmetric nonnegative gate preserves a graph Laplacian. Zero gate slopes recover the static mixture. We call the gated model **MoHL** and the ungated core **MoHL-Static**.

### 3.4 SUPPORT-MATCHED TEST OF GROUP SHARING

**Shared versus free weights.** To isolate the effect of group sharing, we compare MoHL-Static with a freely weighted graph on the same discovered support. Let $A_{\text{shared}}$ be the adjacency induced by the groups. MoHL-Free assigns each supported pair its own weight:

$$A_{\text{free}} = A_{\text{shared}} \odot \frac{\text{softplus}(\vartheta)}{\text{softplus}(0)}, \qquad L_S = \alpha_0 L_{\text{prior}} + \alpha_1 \big(D_{A_{\text{free}}} - A_{\text{free}}\big). \tag{10}$$

At $\vartheta = 0$, the two operators coincide. They use the same support, solver, objective, optimizer, and learning budget; only the parameterization differs. MoHL shares a small number of weights through group structure, whereas MoHL-Free learns each supported coupling independently.

**Sharing trades flexibility for statistical efficiency.** The free model is more expressive, but must estimate many more couplings from the same incomplete observations. Group sharing reduces this estimation burden at the cost of bias when the assumed groups are inaccurate. A stylized regression analysis makes this trade-off explicit: sharing is favored when relational evidence is scarce and group mismatch is small, while free weights become competitive as evidence grows (Proposition 2, Appendix D.6). Section 4.4 tests this prediction directly.

## 4 EXPERIMENTS

We ask three questions: *Does MoHL improve imputation under structured missingness? Are the learned higher-order groups useful? When does group sharing beat free weights?* We test them across three traffic networks, multiple missingness regimes, and support-matched controls.

### 4.1 SETUP

**Data and missingness.** We evaluate PEMS-BAY (speed; 325 sensors), PEMS03 (flow; 358), and PEMS04 (flow; 307), all sampled every five minutes. Native missingness is retained, with zero readings treated as missing on PEMS03/04. The main comparison uses the 100 highest-degree sensors from the largest connected component, the densest seven-day window ($T = 2016$), and $50\%$ injected missingness, with identical masks across methods. We consider three missingness regimes with different surviving evidence. The main results report scattered cell dropout (*Cell*), five nearby sensors missing for 30 minutes (*Cluster*), and a 24-hour gap per sensor (*Block*). Additional experiments include 30-minute and six-hour blocks and a *Mixed* regime combining cell dropout, long gaps, and fully missing sensors.

**Methods and controls.** We compare MoHL with sensor mean, spatial $k$NN, HaLRTC (Liu et al., 2013), LETC (Nie et al., 2023), WDGTC (Li et al., 2020), and three deep imputers: GRIN (Cini et al., 2022), SPIN (Marisca et al., 2022), and ImputeFormer (Nie et al., 2024). Deep models are trained from scratch on the same seven-day window. Our ablation follows the MoHL pipeline: **MoHL-Base** uses the tuned prior graph; subsequent variants add the daily profile, learned relational groups, higher-order interactions (**MoHL-Static**), and the evidence-adaptive gate (**MoHL**). **MoHL-Free** replaces group-shared weights with free weights on the same discovered support. All MoHL variants use matched masks, learning splits, solver, optimizer, and training budget.

### 4.2 PERFORMANCE UNDER STRUCTURED MISSINGNESS

**MoHL excels under long gaps.** Table 1 compares MoHL with five classical and three deep imputers. MoHL achieves the lowest MAE in every reported network and regime, with its largest margins under day-long gaps: $9.2\%$ on PEMS-BAY and $4.5\%$ on PEMS03 over the best baseline. Classical methods degrade sharply as gaps lengthen, while deep imputers remain stronger but are still outperformed by MoHL under Block. MoHL does so with only 57 learned parameters per condition, far fewer than the deep baselines.

**Gains increase as temporal evidence vanishes.** Long gaps remove nearby observations of the missing sensor and force the estimate to rely more on cross-sensor structure. On PEMS-BAY over

Table 1: **MAE on three road networks under three missingness regimes.**

| | PEMS-BAY (mph) | | | PEMS03 (veh/5 min) | | | PEMS04 (veh/5 min) | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Cell | Cluster | Block | Cell | Cluster | Block | Cell | Cluster | Block |
| $k$NN-spatial | 3.450 | 4.428 | 3.469 | 84.08 | 84.50 | 85.89 | 94.46 | 107.7 | 92.41 |
| HaLRTC | 1.116 | 1.818 | 3.203 | 9.800 | 11.55 | 25.11 | 22.95 | 29.18 | 81.98 |
| LETC | 1.748 | 2.147 | 3.178 | 36.79 | 54.49 | 89.63 | 52.33 | 69.26 | 118.7 |
| WDGTC | 2.142 | 3.577 | 13.73 | 13.42 | 14.06 | 40.03 | 26.81 | 31.88 | 103.0 |
| GRIN | 0.706 | 1.291 | 3.029 | 10.36 | 12.61 | 21.57 | 20.26 | 25.97 | 50.03 |
| SPIN | 0.840 | 1.319 | 3.162 | 11.65 | 13.09 | 20.03 | 20.69 | 24.47 | 39.62 |
| ImputeFormer | 0.873 | 1.326 | 3.202 | 10.75 | 11.58 | 14.72 | 20.75 | 23.16 | 31.66 |
| **MoHL** | **0.699** | **1.262** | **2.751** | **9.626** | **11.43** | **14.06** | **19.53** | **23.08** | **31.31** |
| Improvement | +1.0% | +2.2% | +9.2% | +1.8% | +1.0% | +4.5% | +3.6% | +0.3% | +1.1% |

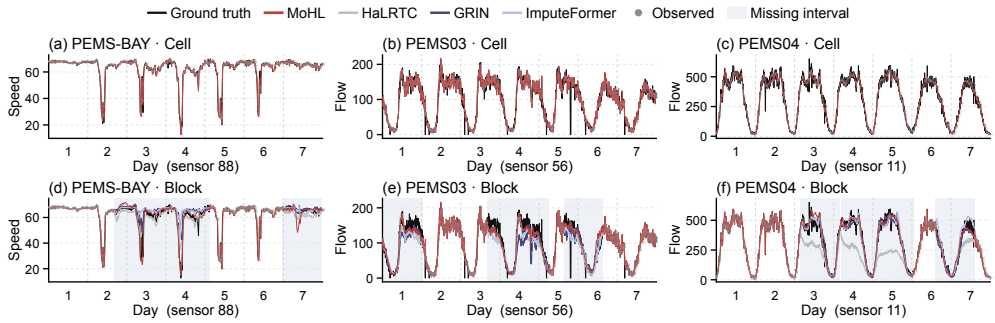


Figure 2: **MoHL imputations over the seven-day window** under Cell (top) and Block (bottom).

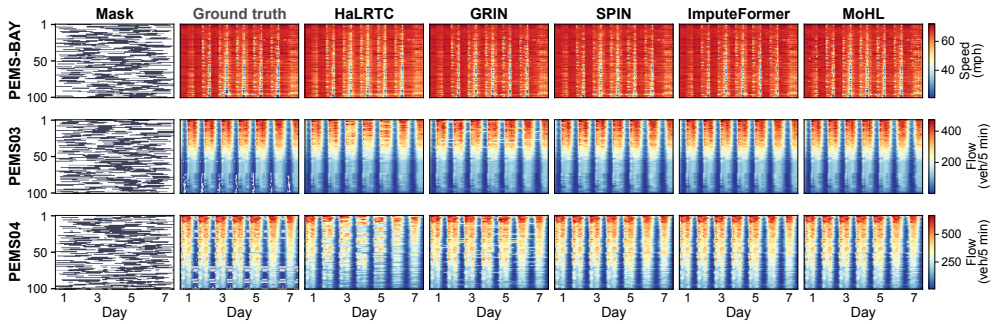


Figure 3: **Reconstructions under Block missingness**: sensor-by-time heatmaps for MoHL and baselines, together with the mask and ground truth.

five seeds, MoHL improves over its profile-plus-prior-graph backbone by only about 4% under Cell, but by 12–39% across six-hour gaps and 11–51% across day-long gaps, depending on distance to the nearest observed value (Table 8). The benefit of learned relational structure therefore increases as local temporal evidence weakens.

**MoHL preserves day-specific dynamics.** Figures 2 and 3 show that MoHL preserves daily cycles and time-varying departures during long gaps. Figure 4 evaluates every day-long imputation against its true day. MoHL attains the highest match on all three networks (43%, 91%, and 80%) while retaining 0.73–0.96 of the ground-truth spread. Its imputations therefore preserve day-specific variation rather than collapsing toward a single typical profile.

**Performance depends on the missingness regime.** Across all six missingness regimes (Appendix B.1), MoHL has the lowest error in most settings, but not all. Under six-hour gaps, all three deep imputers are stronger on PEMS-BAY and ImputeFormer also on PEMS03; under Mixed missingness, SPIN and ImputeFormer are stronger on PEMS04; and HaLRTC is stronger under 30-minute blocks on PEMS03. MoHL's advantage is therefore concentrated where long structured gaps make learned relational structure most valuable.

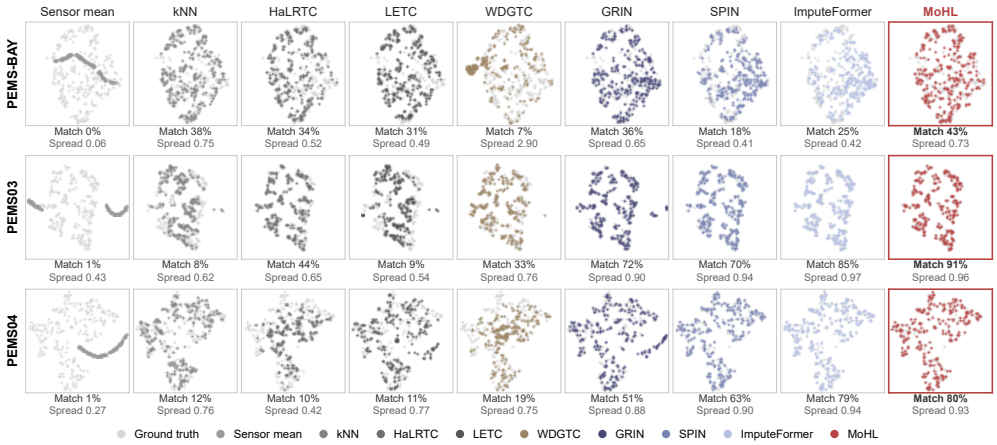


Figure 4: **Imputed days against ground truth**: t-SNE of each day-long gap as ground truth (grey) and as each method's imputation; match and spread are computed in the original space.

Table 2: **Ablation of MoHL**: MAE as components are added, together with the support-matched free-weight control MoHL-Free; best in bold and second best underlined.

| | | PEMS-BAY (mph) | | | PEMS03 (veh/5 min) | | | PEMS04 (veh/5 min) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Description | Cell | Cluster | Block | Cell | Cluster | Block | Cell | Cluster | Block |
| MoHL-Base | prior-graph backbone | 0.732 | 1.273 | 3.128 | 11.11 | 13.31 | 22.98 | 19.82 | 25.17 | 52.71 |
| + profile | adds the daily profile | 0.732 | <u>1.267</u> | 2.882 | 10.99 | 12.97 | 16.07 | 19.63 | 23.61 | 34.57 |
| + pairs | adds groups of order 2 | 0.703 | 1.294 | 2.814 | 9.984 | 12.19 | 15.32 | 19.62 | 23.30 | 32.28 |
| MoHL-Static | adds orders 3–6 | <u>0.700</u> | 1.268 | 2.808 | 9.652 | 11.45 | <u>14.12</u> | <u>19.61</u> | <u>23.14</u> | <u>31.38</u> |
| **MoHL** | adds the gate (full model) | **0.699** | **1.262** | <u>2.751</u> | <u>9.626</u> | <u>11.43</u> | **14.06** | **19.53** | **23.08** | **31.31** |
| MoHL-Free | MoHL-Static, free pair weights | 0.719 | 1.470 | **2.695** | **9.077** | **11.24** | 14.25 | 19.61 | 23.71 | 32.13 |

### 4.3 Learning higher-order structure

**Relational structure drives the gain.** Table 2 adds MoHL's components sequentially. The daily profile provides the strongest first improvement under Block, especially on the flow networks. Selected pairs and higher-order groups then add contemporaneous information that the profile cannot capture, reducing error further on all three networks and most clearly under long gaps. The evidence-adaptive gate provides a smaller final gain. Because the five-seed comparison between MoHL-Static and MoHL-Free is mixed, we evaluate sharing separately over the same five seeds in Section 4.4.

**Higher orders matter under long gaps.** Table 3 adds interaction orders one at a time on PEMS-BAY. Under Cell, pairs are sufficient; under Cluster, order three provides the main gain. Under Block, every order through five improves the estimate, reducing MAE from 2.889 to 2.683 mph (7%). The learned coupling profile shows the same pattern: the relative coupling of the third-ranked neighbor, $\Lambda_4/\Lambda_2$ with $\Lambda_k := \sum_{s\geq k} \alpha_s/\rho_s$, is 0.04 under Cell and 0.09 under 30-minute gaps, but rises to 0.42 under 24-hour and 0.64 under six-hour gaps. The order weights are identifiable on every discovered group set, and the induced coupling profile converges at approximately the $n^{-1/2}$ rate in controlled recovery experiments (Appendix D.3).

Table 3: **Gain by interaction order**.

| | | relational interaction groups, by order | | | | |
|---|---|---|---|---|---|---|
| Regime | | Pairs ($s$=2) | $s$=3 | $s$=4 | $s$=5 | $s$=6 |
| Cell | MAE | **0.740** | 0.741 | 0.741 | 0.741 | 0.741 |
| | Δ | — | +0.001 | +0.001 | +0.001 | +0.001 |
| Cluster | MAE | 1.288 | **1.278** | 1.286 | 1.283 | 1.282 |
| | Δ | — | −0.010 | −0.002 | −0.005 | −0.006 |
| Block | MAE | 2.889 | 2.742 | 2.685 | **2.683** | 2.684 |
| | Δ | — | −0.147 | −0.204 | −0.206 | −0.205 |

**Learned groups generalize to hidden data.** MoHL discovers groups from observed cells only, so we test them on the hidden cells. Under Block missingness, the average hidden-period affinity of the discovered groups exceeds that of equally sized road-graph neighborhoods at every order on PEMS-BAY, decreasing only from 0.63 at order two to 0.55 at order six, compared with 0.52 to 0.48 for the road graph. The same advantage appears at order six on PEMS03 (0.84 versus 0.65)

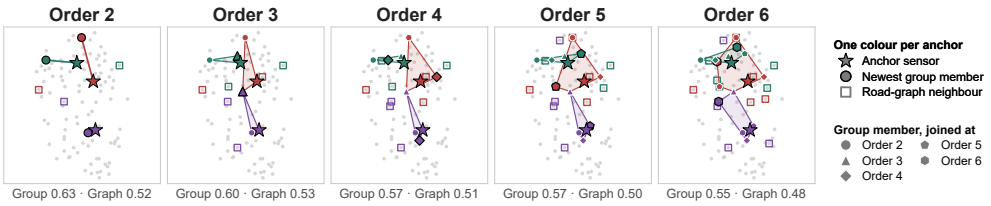


Figure 5: **Discovered groups versus the road graph**: t-SNE of PEMS-BAY sensors using affinity computed on hidden cells, with each anchor's order-$s$ group and its $s-1$ nearest road-graph neighbors.

and PEMS04 (0.69 versus 0.54). These groups can be recovered from incomplete data because their constituent pairs are observed together 3.5–16 times as often as the full groups themselves (Appendix D.2). Thus, pairwise evidence is sufficient to uncover useful higher-order structure.

### 4.4 Evidence-efficient group sharing

**Shared weights outperform free weights on matched support.** MoHL-Static and MoHL-Free use the same sensor pairs, solver, objective, and learning procedure; only the parameterization differs. Over five seeds at $N = 100$ (Figure 6), sharing lowers MAE by 2.7% under Cell, 8.0% under Cluster, and 4.6% under Block. Across all six regimes and both $N = 60$ and $N = 100$, sharing is consistently better, with an average gain of about 4% (Table 7). Because MoHL-Free has the same support but more parameters, the gain comes from sharing statistical evidence rather than additional connectivity.

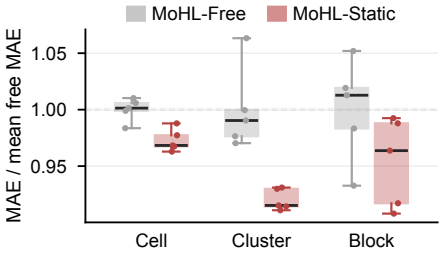


Figure 6: **Shared versus free weights on identical support.**

**Sharing helps most under limited evidence.** Proposition 2 predicts that sharing is most useful when individual couplings are weakly observed. Across observation windows, sharing reduces pooled error by 1.4–2.9% at all but the four-day setting, where the two models are statistically level (Table 9); the predicted crossover thus lies beyond fourteen days, the longest window studied. Controlled experiments likewise show that the excess error of free weights shrinks as joint evidence grows (Appendix D.5). Within a single real network, however, joint evidence varies too little across sensors to reveal a reliable sensor-level trend (Appendix D.6). Together, the results support group sharing as a lower-variance parameterization when relational evidence is limited.

## 5 Conclusion

We introduced **MoHL**, a mixture of hypergraph Laplacians for learning group-wise relational structure from incomplete spatiotemporal data. MoHL discovers overlapping sensor groups, assigns different interaction orders their own Laplacian components, and learns how strongly each order should contribute from mask-matched holdouts. Across three traffic networks, higher-order groups matter most when temporal evidence is weak, and the learned groups capture relationships beyond those encoded by road proximity. On identical relational support, shared weights consistently outperform free weights when relational evidence is limited, showing that the gain comes from parameter sharing rather than added connectivity. MoHL consequently achieves the lowest error among evaluated methods under day-long gaps while using orders of magnitude fewer learned parameters than deep imputers. The key result is that higher-order structure is most useful when missingness makes relational evidence scarce. By sharing statistical strength across groups, MoHL provides a compact and evidence-efficient way to learn cross-sensor dependencies for spatiotemporal imputation.

### AI USE STATEMENT

Generative AI tools were used to assist with writing, coding, debugging, and checks of proofs, notation, and consistency. AI-assisted code and content were reviewed, tested, and verified against the experiments and derivations. The research questions, methods, theoretical claims, and experimental design are the authors' responsibility, and the authors take full responsibility for the final text, claims, results, and artifacts.

### ETHICS STATEMENT

This work uses only public traffic benchmarks (PEMS-BAY, PEMS03 and PEMS04) that record aggregate speeds and flows at fixed road sensors and contain no personal information; no human subjects are involved. Imputation fills in missing sensor readings, so imputed values should be reported downstream as estimates, particularly where they inform traffic management. We are not aware of conflicts of interest or other ethical concerns raised by this work.

### REPRODUCIBILITY STATEMENT

The method is specified in Sections 2–3, with the full estimator in Appendix C.2, the gate and its gradient in Appendix C.4 and all hyperparameters in Appendix C.5. Datasets, windows, sensor selection and missingness regimes are described in Section 4.1. The assumptions and complete proofs of all theoretical results are given in Appendix D, each followed, where its assumptions may not hold on traffic data, by an empirical check. Code is available at `https://anonymous.4open.science/r/MoHL-574F/`.

## REFERENCES

Rie Ando and Tong Zhang. Learning on graph with laplacian regularization. *Advances in neural information processing systems*, 19, 2006.

Alessia Antelmi, Gennaro Cordasco, Mirko Polato, Vittorio Scarano, Carmine Spagnuolo, and Dingqi Yang. A survey on hypergraph representation learning. *ACM Comput. Surv.*, 56(1), August 2023. ISSN 0360-0300. doi: 10.1145/3605776. URL https://doi.org/10.1145/3605776.

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems*, 33:17804–17815, 2020.

Song Bai, Feihu Zhang, and Philip H.S. Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107637. URL https://www.sciencedirect.com/science/article/pii/S0031320320304404.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7: 2399–2434, 2006.

Yoshua Bengio. Gradient-based optimization of hyperparameters. *Neural Computation*, 12(8): 1889–1900, 2000.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/734e6bfcd358e25ac1db0a4241b95651-Paper.pdf.

Xinyu Chen and Lijun Sun. Bayesian temporal factorization for multidimensional time series prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4659–4673, 2022. doi: 10.1109/TPAMI.2021.3066551.

Xinyu Chen, Zhaocheng He, and Jiawei Wang. Spatial-temporal traffic speed patterns discovery and incomplete data recovery via svd-combined tensor decomposition. *Transportation research part C: emerging technologies*, 86:59–77, 2018.

Xinyu Chen, Jinming Yang, and Lijun Sun. A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation. *Transportation Research Part C: Emerging Technologies*, 117:102673, 2020. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2020.102673. URL https://www.sciencedirect.com/science/article/pii/S0968090X2030588X.

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are allset: A multiset function framework for hypergraph neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hpBTIv2uy_E.

Fan RK Chung. The laplacian of a hypergraph. *Expanding graphs*, 10:21–36, 1992.

Andrea Cini, Ivan Marisca, and Cesare Alippi. Filling the G_ap_s: Multivariate time series imputation by graph neural networks. In *International Conference on Learning Representations*, 2022.

Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.119619. URL https://www.sciencedirect.com/science/article/pii/S0957417423001203.

Berihun Fekade, Taras Maksymyuk, Maryan Kyryk, and Minho Jo. Probabilistic recovery of incomplete sensed data in iot. *IEEE Internet of Things Journal*, 5(4):2282–2292, 2017.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3558–3565, Jul. 2019. doi: 10.1609/aaai.v33i01.33013558. URL https://ojs.aaai.org/index.php/AAAI/article/view/4235.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Lawrence A. Klein, Milton K. Mills, and David R. P. Gibson. Traffic Detector Handbook: Third Edition, Volume I. Report FHWA-HRT-06-108, Turner-Fairbank Highway Research Center, Federal Highway Administration, U.S. Department of Transportation, October 2006. URL https://rosap.ntl.bts.gov/view/dot/954.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR '18)*, 2018.

Ziyue Li, Nurettin Dorukhan Sergin, Hao Yan, Chen Zhang, and Fugee Tsung. Tensor completion for weakly-dependent data on graph for metro passenger flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4804–4810, 2020. doi: 10.1609/aaai.v34i04.5915.

Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013. doi: 10.1109/TPAMI.2012.39.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Ivan Marisca, Andrea Cini, and Cesare Alippi. Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 32069–32082. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/cf70320e93c08b39b1b29a348097a376-Paper-Conference.pdf.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, 1982.

Tong Nie, Guoyang Qin, Yunpeng Wang, and Jian Sun. Correlating sparse sensing for large-scale traffic speed estimation: A Laplacian-enhanced low-rank tensor kriging approach. *Transportation Research Part C: Emerging Technologies*, 152:104190, 2023. ISSN 0968-090X. doi: 10.1016/j.trc.2023.104190. URL https://www.sciencedirect.com/science/article/pii/S0968090X23001791.

Tong Nie, Guoyang Qin, Wei Ma, Yuewen Mei, and Jian Sun. ImputeFormer: Low rankness-induced transformers for generalizable spatiotemporal imputation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning*, 2016.

Karelia Pena-Pena, Lucas Taipe, Fuli Wang, Daniel L. Lau, and Gonzalo R. Arce. Learning hypergraphs tensor representations from data via t-hgsp. *IEEE Transactions on Signal and Information Processing over Networks*, 10:17–31, 2024. doi: 10.1109/TSIPN.2023.3345142.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/cfe8504bda37b575c70ee1a8276f3486-Paper.pdf.

Ting Wang, Shengjie Zhao, Wenzhen Jia, and Daqian Shi. A task-oriented spatial graph structure learning method for traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 26(4):4770–4779, 2025.

Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. Inductive graph neural networks for spatiotemporal kriging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5): 4478–4485, May 2021. doi: 10.1609/aaai.v35i5.16575. URL https://ojs.aaai.org/index.php/AAAI/article/view/16575.

Dongwei Xu, Hang Peng, Yufu Tang, and Haifeng Guo. Hierarchical spatio-temporal graph convolutional neural networks for traffic data imputation. *Information Fusion*, 106:102292, 2024.

Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. Hypergcn: A new method for training graph convolutional networks on hypergraphs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1efa39bcaec6f3900149160693694536-Paper.pdf.

Ming Yin, Junbin Gao, and Zhouchen Lin. Laplacian regularized low-rank representation and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):504–517, 2015.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, 2004.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/dff8e9c2ac33381546d96deea9922999-Paper.pdf.

# APPENDIX

## A RELATED WORK

**Spatiotemporal imputation.** PEMS-BAY and METR-LA are the standard test beds (Li et al., 2018). Pooling and interpolation methods exploit no structure beyond a neighborhood mean and remain hard to beat when the assumed structure of richer methods is violated. Low-rank tensor completion (Liu et al., 2013; Chen & Sun, 2022; Chen et al., 2020; Nie et al., 2023) minimizes a nuclear norm over the sensor–time array and is strong under scattered dropout. Graph-aware deep imputers (Cini et al., 2022; Wu et al., 2021; Cao et al., 2018; Du et al., 2023) and generative or attention models (Tashiro et al., 2021; Marisca et al., 2022; Nie et al., 2024) express richer dependence at the cost of per-deployment training; we evaluate GRIN, SPIN and ImputeFormer under our protocol — trained on the same seven-day window, with no external history — and make no claim about them under theirs (Section 4.2). Graph-regularized tensor completion (Li et al., 2020) is the closest prior method in spirit and our most informative classical baseline. A common shortcoming across families is evaluation on one missingness pattern at one rate; we evaluate one model across six regimes that differ in which evidence survives.

**Hypergraph regularization.** Pairwise Laplacian regularization (Ando & Zhang, 2006; Yin et al., 2015) penalizes neighbor differences; the hypergraph Laplacian (Zhou et al., 2006) penalizes within-group variance, and hypergraph neural networks (Feng et al., 2019; Yadati et al., 2019; Bai et al., 2021; Chien et al., 2022) pass messages over hyperedges in the same spirit, typically with the hypergraph given. Learned-hypergraph methods (Pena-Pena et al., 2024; Antelmi et al., 2023) optimize an incidence structure jointly with a task loss under dense observation. Our starting point is the pairwise-expansion identity (Chung, 1992; Zhou et al., 2006) (the "clique expansion" of the hypergraph literature): as a quadratic form, every hypergraph Laplacian is a graph Laplacian in which the pairs of a group share one weight. Prior work has treated this as a representational limitation; we treat it as the mechanism, and build the comparison that isolates it.

**Structured parameter sharing and implicit differentiation.** Sharing parameters across related components to trade bias for variance is an old idea — hierarchical shrinkage, group-structured penalties, weight sharing in convolutional and graph networks (Kipf & Welling, 2016). Our contribution is not the principle but a setting in which it can be measured cleanly: two parameterizations of one operator on one support, learned by one procedure, with the evidence per coupling under experimental control. Hyperparameter gradients through the solution of an optimization problem (Bengio, 2000; Pedregosa, 2016; Lorraine et al., 2020) are usually applied to a handful of scalars; because our estimator is a resolvent, one adjoint solve prices every structural weight, gate parameter and scalar at once, which is what makes learning the structure of a regularizer affordable.

## B ADDITIONAL RESULTS

### B.1 ALL SIX REGIMES

Tables 4–6 extend Table 1 to all six regimes and add every member of the MoHL ladder ($N = 100$, five seeds, masks shared across rows). The columns run from scattered dropout through Cluster and blocks of 30 minutes, six hours and a day (Block-6, Block-72 and Block-288, the suffix giving the block length in five-minute steps; Block-288 is the Block of the main text) to the mixed mask, which adds sensors that never report.

**MoHL's lead is firmest under day-long gaps.** Under Block-288, MoHL has a lower error than every baseline on all three networks, and the ladder rows show the daily profile removing most of the backbone's error and the discovered groups most of the rest. Under six-hour gaps, by contrast, ImputeFormer overtakes MoHL on PEMS-BAY and PEMS03, and on PEMS-BAY GRIN and SPIN do too. Relational structure thus pays most where a sensor's own history is absent for longest; where the holes are shorter, the window still holds enough evidence for densely coupled models.

**Structured masks defeat low-rank completion.** The tensor-completion baselines are the most sensitive to the mask: on PEMS-BAY, HaLRTC and WDGTC end far above even the sensor mean under the mixed mask, and WDGTC also under day-long gaps, whereas the deep imputers and every learned member of the ladder stay well below it. Low-rank completion relies on scattered sampling of the sensor–time array, and missing days and silent sensors remove the entries it needs.

Table 4: **MAE (mph) in all six regimes on PEMS-BAY**; best in bold, second underlined, ranked on unrounded values.

| Method | Cell | Cluster | Block-6 | Block-72 | Block-288 | Mixed |
|---|---|---|---|---|---|---|
| Sensor-mean | 5.044 | 5.124 | 5.023 | 5.221 | 5.251 | 5.136 |
| kNN-spatial | 3.450 | 4.428 | 3.320 | 3.395 | 3.469 | 3.423 |
| HaLRTC | 1.116 | 1.818 | 1.373 | 2.407 | 3.203 | 19.81 |
| LETC | 1.748 | 2.147 | 1.891 | 2.847 | 3.178 | 2.581 |
| WDGTC | 2.142 | 3.577 | 2.241 | 3.291 | 13.73 | 18.21 |
| GRIN | 0.706 | 1.291 | 1.071 | 2.403 | 3.029 | 2.499 |
| SPIN | 0.840 | 1.319 | 1.073 | 2.267 | 3.162 | 2.455 |
| ImputeFormer | 0.873 | 1.326 | 1.119 | **2.207** | 3.202 | 2.674 |
| MoHL-Base | 0.732 | 1.273 | 1.108 | 2.856 | 3.128 | **2.341** |
| Profile + graph | 0.732 | 1.267 | 1.111 | 2.657 | 2.882 | 2.608 |
| + Pairs | 0.703 | 1.294 | 1.071 | 2.519 | 2.814 | 2.520 |
| MoHL-Free | 0.719 | 1.470 | 1.198 | 2.428 | **2.695** | 2.504 |
| MoHL-Static | 0.700 | 1.268 | **1.069** | 2.424 | 2.808 | 2.479 |
| MoHL | **0.699** | **1.262** | 1.069 | 2.461 | 2.751 | 2.453 |

Table 5: **MAE (vehicles per 5 minutes) in all six regimes on PEMS03**; best in bold, second underlined, ranked on unrounded values.

| Method | Cell | Cluster | Block-6 | Block-72 | Block-288 | Mixed |
|---|---|---|---|---|---|---|
| Sensor-mean | 70.03 | 69.55 | 70.32 | 73.29 | 71.02 | 84.57 |
| kNN-spatial | 84.08 | 84.50 | 82.53 | 82.02 | 85.89 | 84.66 |
| HaLRTC | 9.800 | 11.55 | **10.02** | **12.44** | 25.11 | 59.93 |
| LETC | 36.79 | 54.49 | 49.40 | 82.50 | 89.63 | 68.96 |
| WDGTC | 13.42 | 14.06 | 13.50 | 18.46 | 40.03 | 60.89 |
| GRIN | 10.36 | 12.61 | 11.53 | 20.15 | 21.57 | 50.18 |
| SPIN | 11.65 | 13.09 | 12.65 | 17.27 | 20.03 | 43.70 |
| ImputeFormer | 10.75 | 11.58 | 10.91 | 13.31 | 14.72 | 44.78 |
| MoHL-Base | 11.11 | 13.31 | 12.85 | 24.17 | 22.98 | 43.67 |
| Profile + graph | 10.99 | 12.97 | 12.46 | 16.12 | 16.07 | 44.39 |
| + Pairs | 9.984 | 12.19 | 11.37 | 14.23 | 15.32 | 43.02 |
| MoHL-Free | **9.077** | **11.24** | 10.47 | 13.99 | 14.25 | 42.81 |
| MoHL-Static | 9.652 | 11.45 | 10.86 | 13.56 | 14.12 | **42.80** |
| MoHL | 9.626 | 11.43 | 10.77 | 13.61 | **14.06** | 42.81 |

Table 6: **MAE (vehicles per 5 minutes) in all six regimes on PEMS04**; best in bold, second underlined, ranked on unrounded values.

| Method | Cell | Cluster | Block-6 | Block-72 | Block-288 | Mixed |
|---|---|---|---|---|---|---|
| Sensor-mean | 117.6 | 118.6 | 118.5 | 121.2 | 117.9 | 125.9 |
| kNN-spatial | 94.46 | 107.7 | 90.67 | 88.06 | 92.41 | 93.70 |
| HaLRTC | 22.95 | 29.18 | 24.10 | 32.33 | 81.98 | 94.77 |
| LETC | 52.33 | 69.26 | 69.84 | 110.6 | 118.7 | 89.10 |
| WDGTC | 26.81 | 31.88 | 27.51 | 31.77 | 103.0 | 90.58 |
| GRIN | 20.26 | 25.97 | 22.37 | 42.63 | 50.03 | 62.82 |
| SPIN | 20.69 | 24.47 | 22.34 | 33.41 | 39.62 | **51.16** |
| ImputeFormer | 20.75 | 23.16 | 21.83 | 30.17 | 31.66 | 54.84 |
| MoHL-Base | 19.82 | 25.17 | 22.20 | 46.78 | 52.71 | 53.33 |
| Profile + graph | 19.63 | 23.61 | 21.68 | 32.01 | 34.57 | 60.20 |
| + Pairs | 19.62 | 23.30 | 21.57 | 30.11 | 32.28 | 59.03 |
| MoHL-Free | 19.61 | 23.71 | 22.11 | 29.43 | 32.13 | 59.11 |
| MoHL-Static | 19.61 | 23.14 | 21.53 | 29.48 | 31.38 | 58.84 |
| MoHL | **19.53** | **23.08** | **21.47** | **29.26** | **31.31** | 58.72 |

### B.2 Main comparison and ablation as bars

**Both views locate the gains under long gaps.** Figure 7 redraws Table 1 for HaLRTC, the three deep imputers and MoHL, capping and labelling any bar that would leave the axis. Figure 8 redraws Table 2 as each model's MAE increase over the full model, so a longer bar means a larger loss from the components the model lacks. These bars are longest under Block on every network, where

the components of MoHL matter most. The hatched free-weight control falls below zero in some conditions of this five-seed comparison; Appendix B.3 examines sharing across all five seeds.

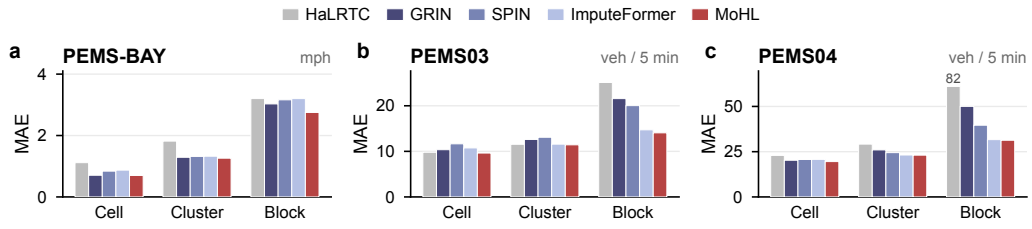


Figure 7: **MoHL against the strongest baselines**: Table 1 as bars.

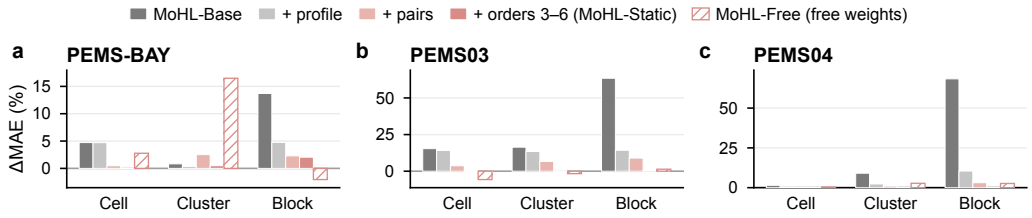


Figure 8: **Ablation of MoHL**: MAE increase of each model of Table 2 over the full model.

### B.3 Shared versus free weights in every regime

**Sharing wins in every regime and at both network sizes.** Table 7 extends Figure 6 to all six regimes and to $N = 60$. Shared weights have the lower error everywhere, by about $4\%$ on average at both sizes; the gain is largest under Cluster and Block-6 and smallest under Mixed. The benefit of sharing thus depends neither on one missingness pattern nor on the size of the network.

Table 7: **Shared versus free weights in all six regimes**: MAE (mph) on PEMS-BAY over five seeds, with the relative change $\Delta$.

| | $N = 60$ | | | $N = 100$ | | |
|---|---|---|---|---|---|---|
| Regime | Free | Shared | $\Delta$ (%) | Free | Shared | $\Delta$ (%) |
| Cell | 0.761 | **0.739** | −2.8 | 0.735 | **0.715** | −2.7 |
| Cluster | 1.429 | **1.290** | −9.7 | 1.368 | **1.259** | −8.0 |
| Block-6 | 1.239 | **1.136** | −8.3 | 1.191 | **1.115** | −6.4 |
| Block-72 | 2.453 | **2.395** | −2.3 | 2.446 | **2.389** | −2.3 |
| Block-288 | 2.725 | **2.672** | −1.9 | 2.688 | **2.563** | −4.6 |
| Mixed | 2.448 | **2.438** | −0.4 | 2.604 | **2.577** | −1.0 |
| Mean $\Delta$ (%) | | −4.2 | | | −4.1 | |

### B.4 Imputations in every regime

**Imputations follow the truth for harder sensors too.** Figure 9 extends Figure 2 to Cluster and to two more sensors per network, at the 25th and 75th percentiles of MoHL's per-sensor error under Block; the median sensor is that of Figure 2, and each sensor keeps its column across regimes. Under Cell and Cluster the imputations stay close to the truth throughout. Inside day-long gaps they keep the daily cycle and, on PEMS-BAY, many of the congestion dips of the day they replace, also for the sensor at the 75th percentile. The accuracy of Table 1 is therefore not carried by a few easy sensors.

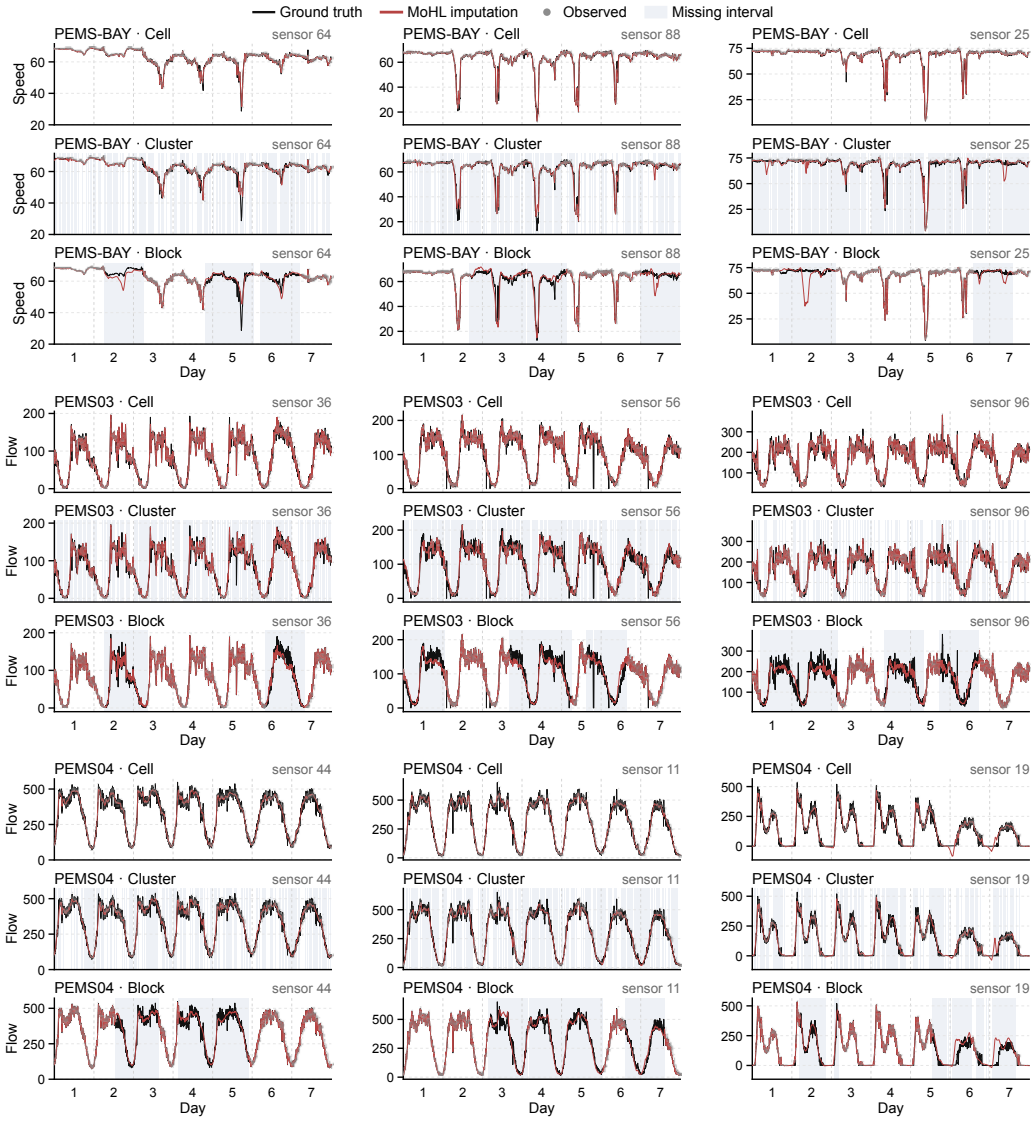


Figure 9: **MoHL's imputations under Cell, Cluster and Block**, three sensors per network.

**The mask leaves no visible imprint on the reconstruction.** Figure 10 shows every sensor at once, with each regime's mask beside MoHL's reconstruction, in the sensor order and color scale of Figure 3. In all three regimes the reconstruction reproduces the daily bands and the differences between sensors in the ground truth, and the blocks of the Block mask cannot be made out in it.

### B.5 Imputed days against the truth

**MoHL imputes the specific day, not a typical one.** Figure 4 embeds every day-long stretch inside a Block gap together with each method's imputation of it, one t-SNE per network, using the stored predictions of Figure 3 (for GRIN and SPIN, from a second training run whose errors are within 3% of Table 1). Two quantities computed in the original space back the picture: the *match*, the share of stretches whose imputation is closer to its own true day than to any other true day, and the *spread*, the total variance of the imputed stretches relative to that of the true ones. MoHL has the highest match on every network with a spread close to the truth, whereas the sensor mean collapses onto one typical day and WDGTC's imputations on PEMS-BAY overshoot into a separate cluster. Removing each sensor's typical day shows where MoHL still falls short: its departures from that day

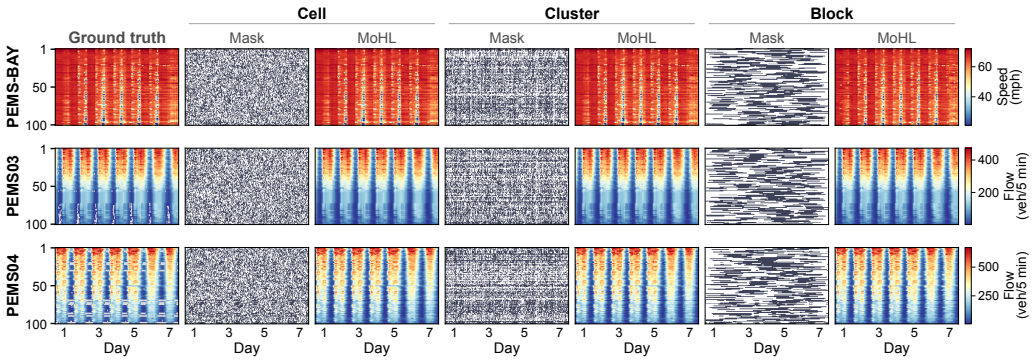


Figure 10: **MoHL's reconstruction of every sensor**: ground truth, mask (dark: missing) and reconstruction under Cell, Cluster and Block.

are compressed relative to those of GRIN and ImputeFormer, because the regularized solve shrinks uncertain departures toward the profile, trading variance for lower error.

### B.6 Temporal gap, window length and interaction orders

**The gain is largest where a sensor's own history is gone.** Table 8 splits the evaluation cells by their temporal gap, the distance to the sensor's nearest observation, and compares MoHL with its profile-plus-prior-graph backbone. Under Cell the gain is small and uniform, under Cluster and Block-6 it grows with the gap, and under Block-72 and Block-288 it is large at every distance. The learned relational structure thus takes over where temporal evidence runs out.

**Sharing wins at every window length studied.** Table 9 varies the window from one to fourteen days. Pooled over regimes, shared weights have the lower error at every window except four days, where the two are level, and the gain does not shrink as the window grows. Even two weeks of data thus leave too little joint evidence for free pair weights to overtake the shared ones.

**The order pattern holds on held-out observed cells.** Table 10 repeats Table 3 on held-out observed cells, with every member fitted on $\Omega_{\text{fit}}$ only. Orders three and four again carry the gain and orders five and six add little, so the choice of orders can be read from observed data alone.

Table 8: **Gain over the backbone by temporal gap**: MAE of profile + prior graph → MoHL on PEMS-BAY ($N = 60$, five seeds).

| | | temporal gap of the evaluation cell (steps) | | | |
|---|---|---|---|---|---|
| Dataset | Regime | 1–2 | 3–12 | 13–72 | $> 72$ |
| PEMS-BAY | Cell | 0.75→0.72 (-4%) | 1.14→1.10 (-4%) | — | — |
| | Cluster | 0.97→0.95 (-2%) | 1.57→1.49 (-5%) | 2.56→2.10 (-18%) | — |
| | Block-6 | 0.94→0.91 (-3%) | 1.38→1.34 (-3%) | 2.15→1.91 (-11%) | — |
| | Block-72 | 1.74→1.06 (-39%) | 2.47→2.01 (-19%) | 3.04→2.54 (-16%) | 2.66→2.33 (-12%) |
| | Block-288 | 2.77→1.35 (-51%) | 2.98→2.29 (-23%) | 2.97→2.60 (-12%) | 2.92→2.60 (-11%) |
| | Mixed | 0.69→0.69 (-1%) | 2.30→1.94 (-16%) | 3.02→2.30 (-24%) | 4.70→4.62 (-2%) |

Table 9: **Shared minus free MAE by window length** on PEMS-BAY ($N = 60$, five seeds), with the pooled relative change and its 95% CI.

| Dataset | Regime | 1 day | 2 days | 4 days | 7 days | 14 days |
|---|---|---|---|---|---|---|
| PEMS-BAY | Cell | $-0.011 \pm 0.002$ | $-0.017 \pm 0.006$ | $-0.025 \pm 0.005$ | $-0.021 \pm 0.005$ | $-0.028 \pm 0.007$ |
| | Block-72 | $-0.003 \pm 0.002$ | $-0.072 \pm 0.021$ | $-0.015 \pm 0.056$ | $-0.041 \pm 0.021$ | $-0.075 \pm 0.031$ |
| | Block-288 | — | $+0.006 \pm 0.016$ | $+0.113 \pm 0.066$ | $-0.114 \pm 0.038$ | $-0.129 \pm 0.021$ |
| | Mixed | $-0.009 \pm 0.005$ | $-0.002 \pm 0.020$ | $+0.017 \pm 0.040$ | $-0.009 \pm 0.008$ | $-0.011 \pm 0.006$ |
| Pooled change (%), 95% CI | | $-1.4$ $(-2.2, -0.7)$ | $-1.5$ $(-2.6, -0.6)$ | $+0.1$ $(-1.7, +2.2)$ | $-2.2$ $(-3.1, -1.3)$ | $-2.9$ $(-4.0, -2.0)$ |

Table 10: **Gain by interaction order on held-out observed cells**, the companion of Table 3.

| Regime | | relational interaction groups, by order | | | | |
|---|---|---|---|---|---|---|
| | | Pairs ($s$=2) | $s$=3 | $s$=4 | $s$=5 | $s$=6 |
| Cell | MAE | 1.041 | **1.040** | **1.040** | **1.040** | **1.040** |
| | Δ | — | −0.001 | −0.001 | −0.001 | −0.001 |
| Cluster | MAE | 1.954 | **1.920** | 1.925 | 1.922 | 1.922 |
| | Δ | — | −0.034 | −0.029 | −0.032 | −0.032 |
| Block-6 | MAE | 1.628 | 1.607 | 1.604 | 1.604 | **1.602** |
| | Δ | — | −0.021 | −0.024 | −0.024 | −0.026 |
| Block-72 | MAE | 3.128 | 2.976 | 2.905 | 2.900 | **2.885** |
| | Δ | — | −0.152 | −0.223 | −0.228 | −0.243 |
| Block-288 | MAE | 3.689 | 3.631 | 3.602 | **3.599** | **3.599** |
| | Δ | — | −0.058 | −0.087 | −0.090 | −0.090 |
| Mixed | MAE | 2.491 | 2.477 | **2.441** | 2.442 | 2.442 |
| | Δ | — | −0.014 | −0.050 | −0.049 | −0.049 |

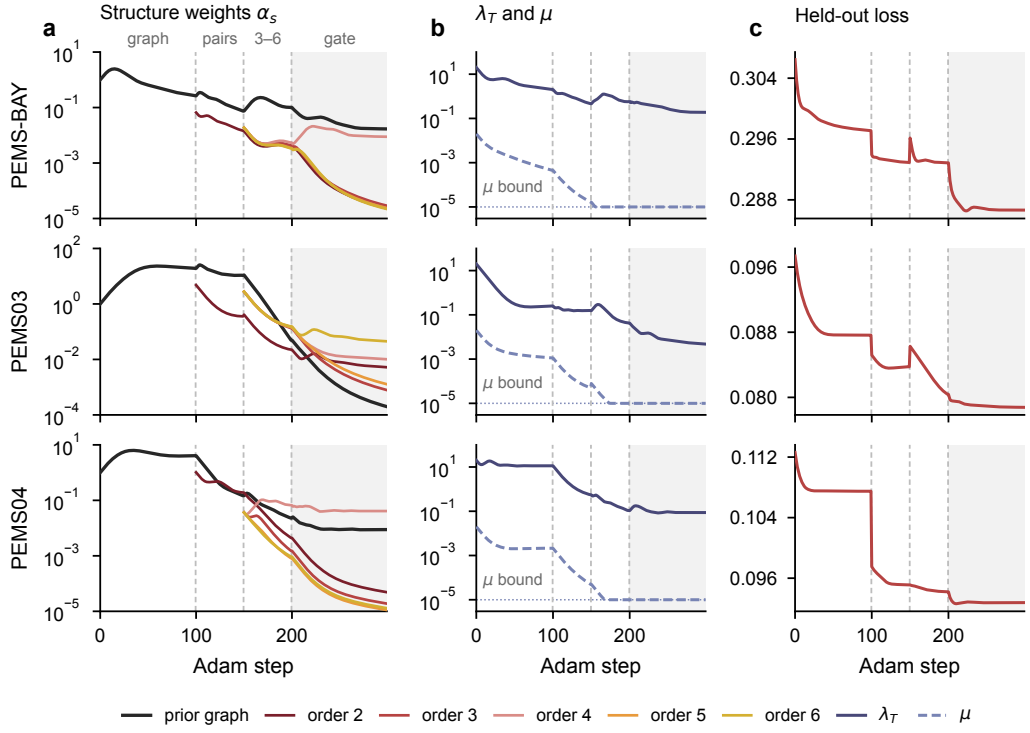


Figure 11: **Learning curves of MoHL under Block**: (a) structure weights $\alpha_s$, (b) $\lambda_T$ and $\mu$, and (c) held-out loss over the staged fit, one row per network; dashed lines separate the stages, the last (shaded) being the gate.

### B.7 Learning curves of the mixture

**The staged fit converges, and each network settles on its own mixture.** Figure 11 follows the parameters of (8) through the fit of MoHL in the Block runs of Table 1, re-run with every step recorded; the final errors reproduce the table. The fit proceeds in warm-started stages: the profile and prior graph, the pairs, orders three to six (MoHL-Static) and the gate (MoHL). The held-out loss falls at every stage, apart from a brief rise when orders three to six enter, which the kept iterate, the one with the lowest objective, does not reflect.

The networks settle differently: on PEMS03 the groups take over and the prior-graph weight falls by two orders of magnitude, on PEMS04 order four dominates, and on PEMS-BAY the prior graph keeps the largest weight. These are single runs of weights that the data identify only coarsely (Appendix D.3), so the curves describe the optimization rather than estimate the weights. On every network the ridge $\mu$ falls to its lower bound: the objective prefers no ridge, and the bound keeps the

solve positive definite (Lemma 4). In the gate stage the $\alpha_s$ are not comparable with the earlier stages, since each edge is also scaled by its gate value.

# C METHOD DETAILS

## C.1 NOTATION

Table 11 collects the symbols used across sections; symbols local to one proof are defined there.

Table 11: **Notation.**

| Symbol | Meaning |
|---|---|
| *Data and missingness* | |
| $N, T$ | number of sensors and of time steps |
| $X^*$ | true values, $N \times T$ |
| $M, \Omega$ | observation mask and the set of observed cells |
| $Y, \xi_{i,t}$ | observed readings $M \odot Y^{\mathrm{obs}}$ and their noise, sub-Gaussian with scale $\sigma$ |
| $\pi_{i,t}, W$ | observation propensity and the weights $W = M \odot \pi^{\odot -1}$ |
| $g_{i,t}$ | temporal gap: distance to the sensor's nearest observation |
| $n_{ij}$ | relational evidence: times sensors $i$ and $j$ are observed together |
| *Graphs and hypergraphs* | |
| $A$ | prior adjacency |
| $D_B$ | degree matrix $\mathrm{diag}(B\mathbf{1})$ of an adjacency $B$ |
| $e, w_e, \mathcal{H}$ | hyperedge (group), its weight, and the hypergraph |
| $P_e$ | hyperedge Laplacian $I_e - \mathbf{1}_e \mathbf{1}_e^\top / \lvert e \rvert$ |
| $A_{\mathcal{H}}$ | adjacency induced by $\mathcal{H}$, with $L_{\mathcal{H}} = D_{A_{\mathcal{H}}} - A_{\mathcal{H}}$ |
| $\mathcal{H}_s$ | retained groups of order $s$ |
| $L_{\mathrm{prior}}, L_{\mathcal{H}_s}$ | $(D_A - A)/\rho_0$ and $\sum_{e \in \mathcal{H}_s} P_e/\rho_s$, each with largest eigenvalue one |
| *Group discovery* | |
| $\tilde{Y}, P$ | standardized observations and the daily profile |
| $\bar{R}$ | one-hour moving average of the residual $(\tilde{Y} - P) \odot M$ |
| $C_{ij}$ | affinity (3) |
| $\mathcal{G}_i^{(s)}, S$ | group of order $s$ around anchor $i$, and the largest order |
| *MoHL* | |
| $\alpha, a_s$ | mixture weights $(\alpha_0, \alpha_2, \ldots, \alpha_S)$ and $a_s = \log \alpha_s$ |
| $L_S(\alpha), A(\alpha)$ | spatial operator (2) and its adjacency (6) |
| $K^{(s)}$ | counts of order-$s$ groups containing each pair |
| $\lambda_T, \mu$ | temporal weight and ridge |
| $\mathcal{A}, \Psi$ | system operator (5) and adjoint state |
| $\Omega_{\mathrm{fit}}, \Omega_{\mathrm{learn}}, \Omega_{\mathrm{sel}}$ | fitting, held-out and selection cells |
| $\ell_\delta, \gamma$ | Huber loss and penalty weight (7) |
| $c_{it}, b_k, h_{it}^k$ | gate context, slopes and values (9) |
| $\Lambda_k$ | coupling profile $\sum_{s>k} \alpha_s/\rho_s$ |
| $A_{\mathrm{free}}, \vartheta$ | free-weight control (10) and its pair logits |

## C.2 THE ESTIMATOR IN FULL

Laplacian-regularized least squares is the workhorse of graph-based semi-supervised learning (Zhou et al., 2004; Belkin et al., 2006; Ando & Zhang, 2006) and the backbone of classical spatiotemporal imputers (Yin et al., 2015; Li et al., 2020); this appendix records the version we run. In the estimator we run, the fidelity term of (4) is $\frac{1}{2}\sum_{(i,t)\in\Omega} W_{i,t}(X_{i,t} - Y_{i,t})^2$ with cell weights $W := M \odot \pi^{\odot -1}$, where $\pi_{i,t}$ is the observation propensity of cell $(i,t)$, estimated from the mask alone as the sensor's smoothed time-of-day occupancy and clipped at $\pi_{\min}$ ($W = M$ recovers the plain estimator (4)); Lemma 1 shows that this makes the fidelity term an unbiased estimate of the squared error over all cells. Setting the gradient of (4) to zero gives the normal equations

$$W \odot X + \lambda_S L_S X + \lambda_T X L_T + \mu X = W \odot Y^{\mathrm{obs}}, \tag{11}$$

whose operator $Z \mapsto W \odot Z + \lambda_S L_S Z + \lambda_T Z L_T + \mu Z$ is symmetric positive definite for $\mu > 0$ (Lemma 4); the solution is unique and is computed by conjugate gradients at $O(\mathrm{nnz}(L_S)\,T + NT)$ per iteration, warm-started and Jacobi-preconditioned. Section 3.2 instantiates (11) with $L_S(\alpha)$ and $\lambda_S$ absorbed into $\alpha$.

**The estimate at a missing cell.** Let $L_S = D_{A(\alpha)} - A(\alpha)$ with row sums $d_i$, and let $L_T$ be the difference Laplacian, $(XL_T)_{i,t} = 2X_{i,t} - X_{i,t-1} - X_{i,t+1}$ for $1 < t < T$. At a missing cell $W_{i,t} = 0$, so the $(i,t)$ entry of (11) is $\lambda_S\big(d_i X_{i,t} - \sum_j A(\alpha)_{ij} X_{j,t}\big) + \lambda_T\big(2X_{i,t} - X_{i,t-1} - X_{i,t+1}\big) + \mu X_{i,t} = 0$, that is,

$$\widehat{X}_{i,t} \;=\; \frac{\lambda_S \sum_j A(\alpha)_{ij}\, \widehat{X}_{j,t} \;+\; \lambda_T\big(\widehat{X}_{i,t-1} + \widehat{X}_{i,t+1}\big)}{\lambda_S d_i + 2\lambda_T + \mu}\,, \tag{12}$$

a weighted average of the coupled detectors at the same time and of the cell's own neighbors in time. Under scattered dropout the temporal neighbors are observed and dominate the average, so the estimate is insensitive to $A(\alpha)$. Inside a long missing interval they are themselves unobserved and supply no information of their own, apart from the observed values at the two ends of the gap, whose influence decays into its interior; for a fully unobserved sensor only the spatial term remains. In these regimes $A(\alpha)_{ij}$ is the weight that detector $j$ receives in the estimate of detector $i$, which is what an entry of the weighted adjacency $A(\alpha)$ does inside the estimator. The identity $\langle X, L_S X\rangle_F = \sum_{i<j} A(\alpha)_{ij} \|X_{i,:} - X_{j,:}\|_2^2$ used there is the standard expansion of a Laplacian quadratic form.

**Lemma 1** (IPW debiasing under MAR)**.** *Assume $M_{i,t}$ is independent of $\xi_{i,t}$ conditional on $X^*$ and $\mathbb{E}[M_{i,t} \mid X^*] = \pi_{i,t} > 0$. For any fixed candidate value $x \in \mathbb{R}$,*

$$\mathbb{E}\left[\frac{M_{i,t}}{\pi_{i,t}}\big(x - Y_{i,t}^{\mathrm{obs}}\big)^2 \,\middle|\, X^*\right] = \mathbb{E}\left[\big(x - X_{i,t}^* - \xi_{i,t}\big)^2 \,\middle|\, X^*\right].$$

*Consequently, the IPW empirical fidelity term is an unbiased estimate of the full-data squared-error fidelity.*

*Proof.* Since $M_{i,t}$ is independent of the observation noise conditional on $X^*$, and $\mathbb{E}[M_{i,t} \mid X^*] = \pi_{i,t}$,

$$\mathbb{E}\left[\frac{M_{i,t}}{\pi_{i,t}}\big(x - Y_{i,t}^{\mathrm{obs}}\big)^2 \,\middle|\, X^*\right] = \frac{1}{\pi_{i,t}}\mathbb{E}[M_{i,t} \mid X^*]\,\mathbb{E}\left[\big(x - X_{i,t}^* - \xi_{i,t}\big)^2 \,\middle|\, X^*\right].$$

The prefactor equals one, giving the result. □

### C.3 Hyperedge Laplacians as shared-weight graphs

Throughout, for $e \subseteq [N]$ with $s = |e| \geq 2$, $P_e := I_e - \frac{1}{s}\mathbf{1}_e\mathbf{1}_e^\top$ is the hyperedge Laplacian of Section 2.2, embedded in $\mathbb{R}^{N\times N}$ and zero outside $e \times e$. This is the convention of the implementation; the size-scaled operator $sP_e$ used in part of the hypergraph literature differs by a constant that the unit-spectral-radius normalization within an order removes.

**Lemma 2** (Per-edge Dirichlet identity)**.** *$P_e$ is symmetric, idempotent ($P_e^2 = P_e$) and positive semidefinite, with $P_e\mathbf{1}_e = 0$ and $\mathrm{rank}(P_e) = s - 1$. For any $X \in \mathbb{R}^{N\times T}$,*

$$\langle X, P_e X\rangle_F \;=\; \sum_{i\in e} \|X_{i,:} - \bar{X}_{e,:}\|_2^2 \;=\; \frac{1}{s}\sum_{\{i,j\}\subseteq e} \|X_{i,:} - X_{j,:}\|_2^2, \qquad \bar{X}_{e,:} := \frac{1}{s}\sum_{i\in e} X_{i,:},$$

*and the form vanishes if and only if all members of $e$ follow the same trajectory.*

*Proof.* On the coordinates of $e$, $\frac{1}{s}\mathbf{1}_e\mathbf{1}_e^\top$ is the orthogonal projector onto $\mathrm{span}(\mathbf{1}_e)$, so $P_e$ is the complementary projector: symmetric, idempotent, positive semidefinite, with kernel $\mathrm{span}(\mathbf{1}_e)$ on $e$ and rank $s-1$. Write $x_i = X_{i,:}$. Then $\langle X, P_e X\rangle_F = \sum_{i\in e}\|x_i\|^2 - \frac{1}{s}\|\sum_{i\in e} x_i\|^2 = \sum_{i\in e}\|x_i\|^2 - s\|\bar{x}_e\|^2 = \sum_{i\in e}\|x_i - \bar{x}_e\|^2$, expanding the square and using $\sum_{i\in e} x_i = s\bar{x}_e$. For the pairwise form, $\sum_{\{i,j\}\subseteq e}\|x_i - x_j\|^2 = \frac{1}{2}\sum_{i,j\in e}\|x_i - x_j\|^2 = s\sum_{i\in e}\|x_i\|^2 - \|\sum_{i\in e} x_i\|^2 = s\,\langle X, P_e X\rangle_F$. A sum of squares vanishes if and only if every $x_i$ equals $\bar{x}_e$. □

**Lemma 3** (Pairwise expansion of the hyperedge Laplacian)**.** *Let $L_{ij} := (e_i - e_j)(e_i - e_j)^\top$ and $A^{(e)} := \frac{1}{s}(\mathbf{1}_e\mathbf{1}_e^\top - I_e)$, the adjacency that places weight $1/s$ on every pair of $e$. Then*

$$P_e \;=\; \frac{1}{s}\sum_{\{i,j\}\subseteq e} L_{ij} \;=\; D_{A^{(e)}} - A^{(e)}.$$

*Consequently every hypergraph Laplacian $L_{\mathcal{H}} = \sum_e w_e P_e$ is the graph Laplacian $D_{A_{\mathcal{H}}} - A_{\mathcal{H}}$ of the weighted adjacency $A_{\mathcal{H}} = \sum_e w_e A^{(e)}$, that is $(A_{\mathcal{H}})_{ij} = \sum_{e \ni i,j} w_e/|e|$ as stated in Section 2.2, and the mixture $L_S(\alpha)$ of* (2) *has the weighted adjacency* (6)*.*

*Proof.* $\sum_{\{i,j\} \subseteq e} L_{ij}$ has diagonal entries $s-1$ on $e$ and off-diagonal entries $-1$ on $e \times e$, so it equals $sI_e - \mathbf{1}_e \mathbf{1}_e^\top = sP_e$; dividing by $s$ gives the first equality. Each vertex of $A^{(e)}$ has degree $(s-1)/s$, so $D_{A^{(e)}} - A^{(e)} = \frac{s-1}{s} I_e - \frac{1}{s}(\mathbf{1}_e \mathbf{1}_e^\top - I_e) = I_e - \frac{1}{s}\mathbf{1}_e \mathbf{1}_e^\top = P_e$. Both $D_{(\cdot)}$ and the identity map are linear in the adjacency, and a nonnegative combination of graph Laplacians is the graph Laplacian of the combined adjacency, which gives the consequence. For (6), the adjacency of $\sum_{e \in \mathcal{H}_s} P_e$ places weight $1/s$ on a pair for each group of $\mathcal{H}_s$ containing it, that is $\frac{1}{s} K^{(s)}$, and $L_{\mathcal{H}_s}$ divides this sum by its spectral radius $\rho_s$. □

### C.4 The evidence-adaptive gate and its gradient

**Matrix form of the gated operator.** With $L_k = D_k - A_k$, $D_k = \mathrm{diag}(\deg_k)$, $\deg_k = A_k \mathbf{1}$, and $H_k = [h_{it}^k]$ the gate values of (9), the gated component acts as $S_k(Z; H_k) = \frac{1}{2}[\, H_k \odot (L_k Z) + (A_k H_k) \odot Z - A_k(H_k \odot Z)\,]$, the temporal term is gated the same way with a weight $\exp(\bar{c}_{it}^\top b_T)$ on each temporal edge, and the operator is $\mathcal{A}Z = W \odot Z + \sum_k \alpha_k S_k(Z; H_k) + \lambda_T Z L_T(g) + \mu Z$, still symmetric positive definite. For every $t$ and every gate value the gated component is a genuine graph Laplacian — symmetric, positive semidefinite, zero row sums — and an edge is strong if either endpoint asks for it. Zero gate slopes reproduce the static operator exactly, so "does the gate pay?" is a nested test; a ridge on the slopes keeps the static model the default.

Since $\langle \Psi, S_k(\hat{X}; H_k) \rangle = \sum_{it} h_{it}^k Q_{it}^k$ with $Q^k = \frac{1}{2}[\Psi \odot \hat{X} \odot \deg_k - \Psi \odot (A_k \hat{X}) - \hat{X} \odot (A_k \Psi) + A_k(\Psi \odot \hat{X})]$, the single adjoint solve of (8) gives $\partial \ell / \partial b_k = -\alpha_k \sum_{it} h_{it}^k Q_{it}^k \, c_{it}$ and the analogous temporal block.

### C.5 Hyperparameters

Table 12 lists the settings shared by all experiments; the window-length study differs only in its step budget.

Table 12: **Hyperparameters.**

| Stage | Setting | Value |
|---|---|---|
| Profile | time-of-day bin | 30 min, smoothed with weights $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ |
| Affinity | moving average | half-width 6 steps, at least 3 observed |
| | shared steps $n_{\min}$ | $\max(30, \min(100, T/10))$ |
| Groups | orders $s$ | 2 to 6 |
| | threshold $c_{\min}$ | 0.2 |
| Learning | loss | Huber, $\delta = 1$ mph |
| | penalty $\gamma$ | $10^{-3}$ (and $10^{-3}$ on the gate slopes) |
| | optimizer | Adam, learning rate 0.1 |
| | steps | 100 profile and prior graph, 50 pairs, 50 orders 3–6, 100 gate; MoHL-Free 100 after the prior graph (same total as shared); window study 100 per model |
| | initialization | $\alpha_0 = 1$, each new order at $\alpha_0/4$, $\lambda_T = 20$, $\mu = 0.02$ (floor $10^{-5}$) |
| Holdout | held-out cells | 0.4 of the observed cells, removed as events shaped like the mask's gaps |
| | held-out sensors | $\mathrm{clip}$(dark fraction, 0.1, 0.3) of observed sensors |
| | split | even day blocks $\to \Omega_{\mathrm{learn}}$, odd $\to \Omega_{\mathrm{sel}}$ |
| Solver | conjugate gradients | tolerance $10^{-7}$; 150 iterations in learning, 600 at deployment; Jacobi preconditioner, warm start |
| Gate | context, clip $\kappa$ | 7 features, $\kappa = 6$ |

# D PROOFS

We use the following conventions: $\langle X, Y\rangle_F = \mathrm{tr}(X^\top Y)$, $e_i$ is the $i$-th standard basis vector of $\mathbb{R}^N$, $X_{i,:}$ is row $i$ of $X$, $\mathbf{1}_e$ and $I_e$ are the all-ones vector and the identity restricted to the coordinates of $e$, $L_{ij} := (e_i - e_j)(e_i - e_j)^\top$ is the Laplacian of a unit pair edge with $\langle X, L_{ij}X\rangle_F = \|X_{i,:} - X_{j,:}\|_2^2$, and $\rho(\cdot)$ is the spectral radius. Spatial operators act by left multiplication and the temporal operator by right multiplication; a linear map $\mathcal{B}$ on $\mathbb{R}^{N\times T}$ is symmetric if $\langle Z, \mathcal{B}Y\rangle_F = \langle \mathcal{B}Z, Y\rangle_F$ and positive definite if in addition $\langle Z, \mathcal{B}Z\rangle_F > 0$ for $Z \neq 0$. Where a result rests on assumptions that traffic data need not meet, its subsection closes with a check on the data and masks of the experiments (PEMS-BAY, with the groups and masks of Table 7 unless stated).

## D.1 WELL-POSEDNESS AND SOLVER COST

The statement, summarized in Section 3.2:

**Lemma 4** (Well-posedness and solver cost)**.** *Let $L_S, L_T \succeq 0$, $\mu > 0$, and $W \geq 0$ entrywise with $w_{\max} := \max_{i,t} W_{i,t}$. The operator $\mathcal{A}Z := W \odot Z + \lambda_S L_S Z + \lambda_T Z L_T + \mu Z$ of* (11) *is symmetric positive definite with spectrum in $[\mu,\ w_{\max} + \lambda_S\rho(L_S) + \lambda_T\rho(L_T) + \mu]$, so* (11) *has a unique solution for every mask, and conjugate gradients started from $X_0$ satisfy*

$$\|X_k - \widehat{X}\|_{\mathcal{A}} \leq 2\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^k \|X_0 - \widehat{X}\|_{\mathcal{A}}, \qquad \kappa \leq \frac{w_{\max} + \lambda_S\rho(L_S) + \lambda_T\rho(L_T) + \mu}{\mu}, \quad (13)$$

*at a cost of $O(\mathrm{nnz}(L_S)\,T + NT)$ operations per iteration.*

*Proof. Step 1: symmetry.* The map $Z \mapsto W \odot Z$ is symmetric because $\langle Y, W \odot Z\rangle_F = \sum_{it} W_{it}Y_{it}Z_{it}$ is symmetric in $(Y, Z)$. For symmetric $L_S$, $\langle Y, L_S Z\rangle_F = \mathrm{tr}(Y^\top L_S Z) = \mathrm{tr}((L_S Y)^\top Z) = \langle L_S Y, Z\rangle_F$. For symmetric $L_T$, $\langle Y, ZL_T\rangle_F = \mathrm{tr}(Y^\top Z L_T) = \mathrm{tr}(L_T Y^\top Z) = \mathrm{tr}((YL_T)^\top Z) = \langle YL_T, Z\rangle_F$. The identity map is symmetric. Hence $\mathcal{A}$, a nonnegative combination of symmetric maps, is symmetric.

*Step 2: lower bound.* For any $Z$,

$$\langle Z, \mathcal{A}Z\rangle_F = \sum_{it} W_{it}Z_{it}^2 + \lambda_S\langle Z, L_S Z\rangle_F + \lambda_T\langle Z, ZL_T\rangle_F + \mu\|Z\|_F^2.$$

The first term is nonnegative because $W \geq 0$. Writing $Z = [z_1, \ldots, z_T]$ by columns, $\langle Z, L_S Z\rangle_F = \sum_t z_t^\top L_S z_t \geq 0$ since $L_S \succeq 0$; writing $Z$ by rows $\zeta_1, \ldots, \zeta_N$, $\langle Z, ZL_T\rangle_F = \sum_i \zeta_i L_T \zeta_i^\top \geq 0$ since $L_T \succeq 0$. Hence $\langle Z, \mathcal{A}Z\rangle_F \geq \mu\|Z\|_F^2$.

*Step 3: upper bound.* By the same decompositions and the Rayleigh quotient, $\sum_t z_t^\top L_S z_t \leq \rho(L_S)\|Z\|_F^2$ and $\sum_i \zeta_i L_T \zeta_i^\top \leq \rho(L_T)\|Z\|_F^2$, while $\sum_{it} W_{it}Z_{it}^2 \leq w_{\max}\|Z\|_F^2$. Therefore

$$\mu\|Z\|_F^2 \leq \langle Z, \mathcal{A}Z\rangle_F \leq \big(w_{\max} + \lambda_S\rho(L_S) + \lambda_T\rho(L_T) + \mu\big)\|Z\|_F^2.$$

*Step 4: spectrum, uniqueness, conditioning.* Identify $\mathbb{R}^{N\times T}$ with $\mathbb{R}^{NT}$; $\mathcal{A}$ is then a symmetric matrix whose Rayleigh quotient lies in the interval of Step 3, so all of its eigenvalues do. In particular $\mathcal{A} \succ 0$, it is invertible, and (11) has the unique solution $\widehat{X} = \mathcal{A}^{-1}(W \odot Y^{\mathrm{obs}})$. Because (4) is a strictly convex quadratic with gradient $\mathcal{A}X - W \odot Y^{\mathrm{obs}}$, this solution is its unique minimizer. The condition number satisfies $\kappa = \lambda_{\max}(\mathcal{A})/\lambda_{\min}(\mathcal{A}) \leq (w_{\max} + \lambda_S\rho(L_S) + \lambda_T\rho(L_T) + \mu)/\mu$.

*Step 5: conjugate gradients.* For a symmetric positive-definite system with condition number $\kappa$, the $k$-th conjugate-gradient iterate minimizes the $\mathcal{A}$-norm of the error over the $k$-th Krylov subspace, and the classical Chebyshev-polynomial argument yields $\|X_k - \widehat{X}\|_{\mathcal{A}} \leq 2\big((\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)\big)^k \|X_0 - \widehat{X}\|_{\mathcal{A}}$. Consequently $k \geq \frac{1}{2}\sqrt{\kappa}\,\ln(2/\varepsilon)$ iterations suffice for a relative error of $\varepsilon$ in the $\mathcal{A}$-norm.

*Step 6: cost per iteration.* One application of $\mathcal{A}$ consists of an entrywise product ($NT$ operations), a sparse left multiplication by $L_S$ ($\mathrm{nnz}(L_S)\,T$ operations), a right multiplication by the tridiagonal $L_T$ ($O(NT)$ operations), and a scaling ($NT$ operations), for a total of $O(\mathrm{nnz}(L_S)\,T + NT)$. □

**Spectral radii of the MoHL operator.** The difference temporal Laplacian is the graph Laplacian of the path on $T$ vertices, whose eigenvalues are $2 - 2\cos(\pi k/T)$, $k = 0, \dots, T-1$; hence $\rho(L_T) < 4$. For the MoHL operator (5), in which $\lambda_S$ is absorbed into the weights, the triangle inequality for the spectral norm and $\rho(L_{\text{prior}}) = \rho(L_{\mathcal{H}_s}) = 1$ give $\rho(L_S(\alpha)) \le \sum_k \alpha_k$, so

$$\kappa \le \frac{w_{\max} + \sum_k \alpha_k + 4\lambda_T + \mu}{\mu}.$$

The bound makes the cost of the learned operating point explicit: learning drives $\mu$ down, which improves accuracy and worsens conditioning in the same motion, and it is $\mu$, not the size of the Laplacian dictionary, that governs the number of iterations. The IPW weights are bounded by $w_{\max} \le 1/\pi_{\min}$, since the propensities are clipped at $\pi_{\min}$ (Appendix C.2).

### D.2 Pairwise versus group observation rate

The statement, summarized in Section 3.1:

**Lemma 5** (Pairwise versus group observation rate)**.** *Under an i.i.d. Bernoulli observation model with rate $\pi$, the number of jointly observed samples for a fixed pair is $n_{ij} \sim \text{Binomial}(T, \pi^2)$, whereas for a fixed group of size $s$ it is $n_e \sim \text{Binomial}(T, \pi^s)$. For any $0 < \delta < 1$, $\Pr[n_{ij} \le (1-\delta)T\pi^2] \le e^{-\delta^2 T\pi^2/2}$ and $\Pr[n_e \le (1-\delta)T\pi^s] \le e^{-\delta^2 T\pi^s/2}$. The effective sample sizes are therefore of order $T\pi^2$ and $T\pi^s$, and their ratio is $\pi^{-(s-2)}$.*

*Proof.* Suppose the mask entries $M_{it}$ are independent Bernoulli variables with $\Pr[M_{it} = 1] = \pi \in (0, 1]$.

*Pair counts.* For a fixed pair $(i, j)$ with $i \ne j$ define $B_t = M_{it}M_{jt} \in \{0, 1\}$. Since $M_{it}$ and $M_{jt}$ are independent, $\Pr[B_t = 1] = \pi^2$; and since $B_t$ depends only on the mask entries in column $t$, the variables $B_1, \dots, B_T$ are independent. Therefore

$$n_{ij} = \sum_{t=1}^{T} B_t \sim \text{Binomial}(T, \pi^2), \qquad \mathbb{E}n_{ij} = T\pi^2.$$

*Group counts.* For a fixed group $e$ of size $s$ define $G_t = \prod_{i \in e} M_{it}$. By independence across sensors $\Pr[G_t = 1] = \pi^s$, and by independence across columns the $G_t$ are independent, so

$$n_e = \sum_{t=1}^{T} G_t \sim \text{Binomial}(T, \pi^s), \qquad \mathbb{E}n_e = T\pi^s.$$

*Concentration.* Let $n$ be a sum of independent Bernoulli variables with $\mathbb{E}n = \nu$. The multiplicative Chernoff bounds state that for $0 < \delta < 1$,

$$\Pr[n \le (1-\delta)\nu] \le \exp\bigl(-\delta^2\nu/2\bigr), \qquad \Pr[n \ge (1+\delta)\nu] \le \exp\bigl(-\delta^2\nu/3\bigr).$$

Applying the lower-tail bound with $\nu = T\pi^2$ and with $\nu = T\pi^s$ gives the two inequalities of the lemma. In particular, with probability at least $1 - e^{-\delta^2 T\pi^2/2}$ a pair retains at least $(1-\delta)T\pi^2$ jointly observed samples, while the same guarantee for a group requires $T\pi^s$ in the exponent.

*Ratio.* The expected counts satisfy $\mathbb{E}n_{ij}/\mathbb{E}n_e = \pi^2/\pi^s = \pi^{-(s-2)}$. Equivalently, holding the expected number of usable samples fixed at $n_0$ requires a window of length $T = n_0\pi^{-2}$ for pairwise statistics and $T = n_0\pi^{-s}$ for statistics that need all $s$ members simultaneously. □

**Numerical illustration.** At $\pi = 0.5$ and $T = 2016$, a pair is jointly observed $504$ times in expectation, a triple $252$ times, and a five-sensor group $63$ times; at $T = 576$ the corresponding numbers are $144$, $72$, and $18$. A statistic that needs a five-sensor group to be fully observed thus has an effective sample size below $20$ in the two-day setting, whereas every pair inside it still has more than a hundred.

**Heterogeneous rates.** If $\Pr[M_{it} = 1] = \pi_{it}$ varies across cells, the same argument gives $\mathbb{E}n_{ij} = \sum_t \pi_{it}\pi_{jt}$ and $\mathbb{E}n_e = \sum_t \prod_{l \in e} \pi_{lt}$, and the Chernoff bounds hold verbatim because they require only independence, not identical distribution. Since $\prod_{l \in e} \pi_{lt} \le \pi_{it}\pi_{jt}$ for every pair $\{i, j\} \subseteq e$, the group count is stochastically dominated by every pair count inside it.

**Scope.** The experimental masks contain clustered missing intervals, temporal blocks, and fully unobserved sensors, so their cells are not independent Bernoulli draws. The lemma is used only to explain the scaling gap between pairwise and all-member co-observation. The dominance $n_e \le n_{ij}$ for $\{i, j\} \subseteq e$ holds deterministically for *any* mask, since a timestep at which all members are observed is in particular one at which $i$ and $j$ are; the MoHL construction relies on nothing stronger.

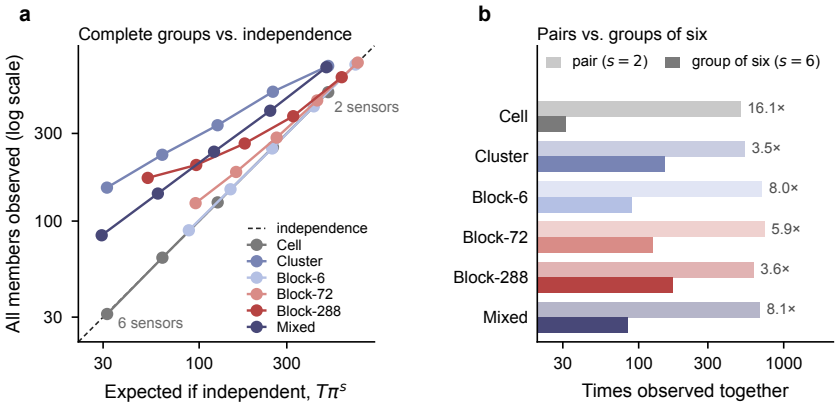


Figure 12: **Why affinities are pairwise (Lemma 5)**: (a) how often discovered groups are fully observed, against independent missingness, and (b) pairs against groups of six, on each regime's mask.

**Pairs carry several times the evidence of whole groups.** Lemma 5 assumes independent missingness, under which a pair is observed together far more often than a whole group. Real gaps cluster in space and time, which favors whole groups, and indeed groups are fully observed more often than independence predicts under the clustered, long-gap and mixed masks (Figure 12a). Their pairs are nonetheless observed together 3.5 to 16 times as often as groups of six on every mask (Figure 12b). Pairwise affinities thus remain the better-supported basis for discovering groups, even under strongly dependent missingness.

### D.3 Proof of Theorem 1

Write $\mathcal{H}_s$ for the set of retained order-$s$ groups (Section 3.1), $L_s := \sum_{e \in \mathcal{H}_s} P_e$ for their unnormalized aggregate, so that the order-$s$ component of (2) is $L_{\mathcal{H}_s} = L_s/\rho_s$ with $\rho_s := \rho(L_s)$, and $L_{\text{prior}} = (D_A - A)/\rho_0$ with $\rho_0 := \rho(D_A - A)$. For $i \ne j$ and an order $s$ let

$$K_{ij}^{(s)} = \big|\{e \in \mathcal{H}_s : \{i, j\} \subseteq e\}\big|$$

be the number of order-$s$ groups that contain both sensors, let $K^{(s)}$ be the symmetric matrix of these counts with zero diagonal, and let $E = \{\{i, j\} : K_{ij}^{(s)} > 0 \text{ for some } s\}$ be the relational support, with $m = |E|$; let $\Phi$ be the matrix whose rows are the profiles $\varphi(i, j)$ of all pairs with a nonzero profile (the pairs of $E$ and the prior edges), and abbreviate the adjacency of (6) as $A(\alpha)$. An empty order has no component, so $\alpha$ ranges over the nonempty orders.

*Proof of Theorem 1. Part (i).* By Lemma 3, $L_s = \frac{1}{s} \sum_{e \in \mathcal{H}_s} \sum_{\{i,j\} \subseteq e} L_{ij}$, every group of $\mathcal{H}_s$ having size $s$. Exchanging the two sums, the pair Laplacian $L_{ij}$ appears once for every group of $\mathcal{H}_s$ that contains $\{i, j\}$, so $L_s = \frac{1}{s} \sum_{i<j} K_{ij}^{(s)} L_{ij}$. By the identity $D_B - B = \sum_{i<j} B_{ij} L_{ij}$ for a symmetric nonnegative $B$ with zero diagonal (which holds entrywise: the diagonal of both sides is the row sum

of $B$, the off-diagonal entries are $-B_{ij}$), this equals $\frac{1}{s}(D_{K^{(s)}} - K^{(s)})$. Substituting into (2), using $L_{\text{prior}} = (D_A - A)/\rho_0$ and the linearity of $B \mapsto D_B - B$, gives (6).

*Part (ii).* By (6), $A(\alpha)_{ij} = \langle \beta, \varphi(i,j) \rangle$ with $\beta = (\alpha_0/\rho_0, \alpha_2/(2\rho_2), \ldots, \alpha_S/(S\rho_S))$. Equal profiles therefore give equal weights for every $\alpha$. Conversely, if two profiles differ in some coordinate, choosing $\alpha$ supported on that coordinate gives the two pairs different weights.

*Part (iii).* Order the columns of $\Phi$ as $(S, S-1, \ldots, 2, 0)$, where 0 is the prior column, and select the rows $(p_S, p_{S-1}, \ldots, p_2, p_0)$, where $p_s$ is the pair that first appears at order $s$ and $p_0$ is the prior edge contained in no group, omitting empty orders. Row $p_s$ has a positive entry in column $s$, zeros in the columns $s' < s$ by the first-appearance condition, and a zero in the prior column because $p_s$ is off the prior graph; row $p_0$ is zero except for a positive entry in the prior column. The selected square submatrix is therefore lower triangular with a positive diagonal, hence invertible, so $\Phi$ has full column rank. Then $A(\alpha)$ determines $\beta$ and, since every $\rho_k > 0$ is known, determines $\alpha$. □

**Proposition 1** (Cone structure of the shared family)**.** *The shared family* $\{A(\alpha) : \alpha \geq 0\}$ *is a polyhedral cone of dimension* $\operatorname{rank}(\Phi) \leq S$*, and its relational part is contained in the closure of the free family* $\{A_{\text{shared}} \odot c : c \in (0,\infty)^E\}$ *of* (10)*, which has dimension* $m$*.*

*Proof.* The map $\beta \mapsto \Phi\beta$ is linear from $\mathbb{R}^S$ to the space of pair weights, and $\beta$ ranges over the nonnegative orthant as $\alpha$ does, because each $\rho_k > 0$. The image of an orthant under a linear map is a polyhedral cone whose dimension is the rank of the map. For the containment, let $B = \sum_{s \geq 2} \beta_s K^{(s)}$ be the relational part of a shared adjacency. It is supported on $E$, and the initial shared adjacency $A_{\text{shared}}$ is strictly positive on $E$, since it is of the same form with all coefficients positive. Hence $B = A_{\text{shared}} \odot c$ with $c_{ij} = B_{ij}/(A_{\text{shared}})_{ij} \geq 0$ on $E$, and $B$ is the limit of $A_{\text{shared}} \odot c^{(r)}$ with $c^{(r)} = c + 1/r > 0$. The free family is the image of $(0,\infty)^E$ under an injective linear map, so it has dimension $m$. □

**Reading the theorem.** Part (ii) is the precise meaning of structured weight sharing for the operator that MoHL deploys: a free graph on the relational support has one parameter per pair, whereas MoHL has one per co-membership coordinate, and all pairs with the same co-membership profile share a weight. For a single chain, $K^{(s)}_{v_a v_b} = \mathbf{1}[s \geq \max(a,b)]$ and (6) reduces to Corollary 2(a). Proposition 1 supplies the inclusion used in Corollary 4, and part (iii) is what licenses reading the learned coupling profile in Section 4.3.

**The data determine the coupling profile, not each single weight.** Section 4.3 reads the learned order weights, which is meaningful only if the data determine them. The condition of Theorem 1(iii) holds for every discovered group set, with a well-conditioned profile matrix (Figure 13a). To measure how much data recovery takes, we fix known weights on the Cell groups, draw data from the Gaussian model of Theorem 2(b) under a Cell mask, and learn the weights back over windows of 1.75 to 112 days. Single order weights converge slowly, because the co-membership of orders four to six is nearly collinear, whereas the coupling profile $\Lambda_k$, which pools the orders, converges at about the $n^{-1/2}$ rate (Figure 13b). When two orders are made to share their groups, their weights stop converging while their sum still does (Figure 13c). The coupling profile, not single weights, is therefore what a learned MoHL can be read for; on the real data it extends further down each sensor's neighbor ranking as gaps lengthen (Figure 13d).

### D.4 The chain spectrum: shared weights are a monotone coupling profile

On one anchor's chain — the unit Definition 1 builds — the spectrum of the mixture is exact: nested sets give nested projectors, $P_e P_{e'} = P_{e'} P_e = P_e$ for $e \subseteq e'$, so one basis diagonalizes every mixture at once.

**Corollary 1** (Shared weights are a monotone coupling profile)**.** *Let* $e_2 \subset \cdots \subset e_S$ *be any sets with* $|e_k| = k$*, written* $e_k = \{v_1, \ldots, v_k\}$*; let* $L_{\text{ch}}(\alpha) := \sum_{s=2}^{S} \alpha_s P_{e_s}$ *with* $\alpha \geq 0$*, and let* $h_k := \left(\frac{1}{k-1} \sum_{j<k} e_{v_j}\right) - e_{v_k}$ *compare the k-th member against the mean of the earlier ones (*$e_v$ *the standard basis vector). Then:*

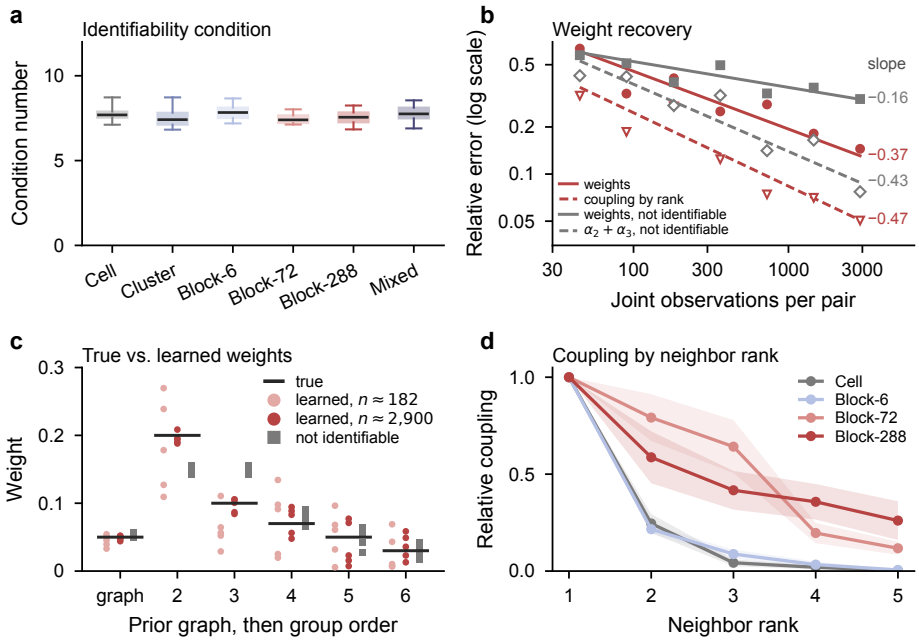


Figure 13: **Learning the order weights (Theorem 1(iii))**: (a) conditioning of the profile matrix, (b, c) recovery of known weights with a non-identifiable control, and (d) the learned coupling of each neighbor relative to the first.

*(i)* ***One basis for the whole family.*** *For every $\alpha \geq 0$, $\{\mathbf{1}_{e_S}, h_2, \ldots, h_S\}$ is an orthogonal eigenbasis of $L_{\mathrm{ch}}(\alpha)$, with $L_{\mathrm{ch}}(\alpha)\mathbf{1}_{e_S} = 0$ and*

$$L_{\mathrm{ch}}(\alpha)\, h_k \;=\; \Lambda_k\, h_k, \qquad \Lambda_k := \sum_{s=k}^{S} \alpha_s. \tag{14}$$

*(ii)* ***Exact range.*** *The family is precisely the operators that are diagonal in this basis with a nonincreasing nonnegative spectrum,*

$$\big\{L_{\mathrm{ch}}(\alpha) : \alpha \geq 0\big\} \;=\; \Big\{ \textstyle\sum_{k=2}^{S} \Lambda_k \frac{h_k h_k^\top}{\|h_k\|_2^2} \;:\; \Lambda_2 \geq \cdots \geq \Lambda_S \geq 0 \Big\},$$

*and the correspondence is a bijection with $\alpha_s = \Lambda_s - \Lambda_{s+1}$ ($\Lambda_{S+1} := 0$): the shared weights are exactly the increments of a coupling profile nonincreasing in neighbor rank.*

*Proof. Nested projectors.* Let $e \subseteq e'$ with $|e| = s$ and $|e'| = s'$. The inclusion gives $I_e I_{e'} = I_e$, $I_e \mathbf{1}_{e'} = \mathbf{1}_e$ and $\mathbf{1}_e^\top \mathbf{1}_{e'} = s$, so $P_e \mathbf{1}_{e'} = \mathbf{1}_e - \frac{1}{s}\mathbf{1}_e\, s = 0$ and hence

$$P_e P_{e'} = P_e I_{e'} - \tfrac{1}{s'}\,(P_e \mathbf{1}_{e'})\mathbf{1}_{e'}^\top = P_e.$$

The right-hand side is symmetric, so $P_{e'} P_e = (P_e P_{e'})^\top = P_e = P_e P_{e'}$. Any two members of the chain $e_2 \subset \cdots \subset e_S$ are nested, so the components of $L_{\mathrm{ch}}(\alpha)$ commute pairwise.

*Part (i): the contrasts are orthogonal.* Each $h_k$ is supported on $e_k \subseteq e_S$ and satisfies $\mathbf{1}^\top h_k = (k-1)\cdot \frac{1}{k-1} - 1 = 0$, so $h_k \perp \mathbf{1}_{e_S}$. For $k < l$ the support of $h_k$ lies in $e_k \subseteq e_{l-1}$, on which $h_l$ is constant and equal to $1/(l-1)$; hence $\langle h_k, h_l \rangle = \frac{1}{l-1}\,\mathbf{1}^\top h_k = 0$. Finally $\|h_k\|_2^2 = (k-1)\cdot\frac{1}{(k-1)^2} + 1 = \frac{k}{k-1} > 0$. Thus $\{\mathbf{1}_{e_S}, h_2, \ldots, h_S\}$ consists of $S$ nonzero, mutually orthogonal vectors in the $S$-dimensional coordinate subspace of $e_S$, and is an orthogonal basis of it.

*Part (i): action of one component.* Fix $s$ and $k$. If $s \geq k$, then $\mathrm{supp}(h_k) \subseteq e_k \subseteq e_s$, so $I_{e_s} h_k = h_k$ and $\mathbf{1}_{e_s}^\top h_k = \mathbf{1}^\top h_k = 0$, giving $P_{e_s} h_k = h_k$. If $s < k$, then $e_s \subseteq e_{k-1}$, on which $h_k$ equals the constant $1/(k-1)$; hence $I_{e_s} h_k = \frac{1}{k-1}\mathbf{1}_{e_s}$ and $\mathbf{1}_{e_s}^\top h_k = \frac{s}{k-1}$, giving $P_{e_s} h_k = \frac{1}{k-1}\mathbf{1}_{e_s} - \frac{1}{s}\cdot$

$\frac{s}{k-1}\mathbf{1}_{e_s} = 0$. Summing over $s$,

$$L_{\mathrm{ch}}(\alpha)\, h_k = \sum_{s=k}^{S} \alpha_s\, h_k = \Lambda_k h_k.$$

Also $P_{e_s}\mathbf{1}_{e_S} = \mathbf{1}_{e_s} - \frac{1}{s}\mathbf{1}_{e_s}(\mathbf{1}_{e_s}^\top \mathbf{1}_{e_S}) = 0$ for every $s$, so $L_{\mathrm{ch}}(\alpha)\mathbf{1}_{e_S} = 0$, and $L_{\mathrm{ch}}(\alpha)$ vanishes on every coordinate outside $e_S$. Since $\alpha_s \geq 0$, $\Lambda_k - \Lambda_{k+1} = \alpha_k \geq 0$ and $\Lambda_S = \alpha_S \geq 0$, which is the ordering of part (ii).

*Part (ii).* The relations $\Lambda_S = \alpha_S$ and $\Lambda_k - \Lambda_{k+1} = \alpha_k$ for $k < S$ define a triangular linear system with unit diagonal, so $\alpha \mapsto \Lambda$ is a linear bijection of $\mathbb{R}^{S-1}$ with the stated inverse. Under it, $\alpha \geq 0$ holds if and only if $\Lambda_S \geq 0$ and $\Lambda_k \geq \Lambda_{k+1}$ for all $k < S$, which is the isotonic cone. Because a symmetric matrix is determined by its eigenpairs,

$$\big\{L_{\mathrm{ch}}(\alpha) : \alpha \geq 0\big\} = \Big\{\sum_{k=2}^{S} \Lambda_k \frac{h_k h_k^\top}{\|h_k\|_2^2} : \Lambda_2 \geq \cdots \geq \Lambda_S \geq 0\Big\},$$

the set of operators that are diagonal in the Helmert basis with a nonincreasing nonnegative spectrum. □

**Scope.** The corollary is exact for a single chain and, because the corresponding matrices are block diagonal, for any collection of chains with pairwise disjoint supports. In MoHL the chains of different anchors overlap, each order is normalized by $\rho(L_s)$ (which only rescales $\alpha_s$), and one $\alpha_s$ is shared by all anchors. Overlapping chains are in general not nested, so their Laplacians need not commute and the full operator has no closed-form eigenbasis. Corollary 1 is therefore a statement about the local structure that every anchor's chain contributes; Theorem 1 is the exact statement for the full operator, and Corollary 2(a) is its single-chain special case.

### D.5 Proof of Theorem 2

*Proof. Part (a).* Let $\bar{x}_e = \frac{1}{s}\sum_{i\in e} x_i$. For every $m \in \mathbb{R}$, $\sum_{i\in e}(x_i - m)^2 = \sum_{i\in e}(x_i - \bar{x}_e)^2 + s(m - \bar{x}_e)^2$, since the cross term $2(\bar{x}_e - m)\sum_{i\in e}(x_i - \bar{x}_e)$ vanishes. With the Gaussian integral $\int_{\mathbb{R}} e^{-\alpha s(m-\bar{x}_e)^2/2}\, dm = (2\pi/(\alpha s))^{1/2}$,

$$\int_{\mathbb{R}} \exp\Big(-\frac{\alpha}{2}\sum_{i\in e}(x_i - m)^2\Big) dm = \Big(\frac{2\pi}{\alpha s}\Big)^{1/2} \exp\Big(-\frac{\alpha}{2}\sum_{i\in e}(x_i - \bar{x}_e)^2\Big).$$

By Lemma 2 applied to a single column, $x^\top P_e x = \sum_{i\in e}(x_i - \bar{x}_e)^2$, so the exponential on the right equals $\exp(-\frac{\alpha}{2}x^\top P_e x)$, and rearranging gives Theorem 2(a). The integrand is, up to a constant, the joint density of $(x_i)_{i\in e}$ given the level $m$ under $x_i = m + \varepsilon_i$ with independent $\varepsilon_i \sim \mathcal{N}(0, 1/\alpha)$; integrating against the flat prior on $m$ gives the marginal.

*Part (b).* Let $\mathcal{P}Z = L_S(\alpha^\circ)Z + \lambda_T Z L_T + \mu Z$. By Lemma 4 and $\mu > 0$, $\mathcal{P}$ is symmetric positive definite, so $X \sim \mathcal{N}(0, \mathcal{P}^{-1})$ is a proper prior with density proportional to $\exp(-\frac{1}{2}\langle X, \mathcal{P}X\rangle_F)$. With $Y$ standing for the residual data $\tilde{Y} - P$ of (5), the likelihood of the observed cells is proportional to $\exp(-\frac{1}{2}\sum_{it} W_{it}(X_{it} - Y_{it})^2)$ with $W_{it} = 0$ on unobserved cells. The log posterior is therefore, up to a constant,

$$-\frac{1}{2}\langle X, W \odot X\rangle_F - \frac{1}{2}\langle X, \mathcal{P}X\rangle_F + \langle X, W \odot Y\rangle_F = -\frac{1}{2}\langle X, \mathcal{A}X\rangle_F + \langle X, W \odot Y\rangle_F,$$

a Gaussian log density with precision $\mathcal{A}$ and mean $\mathcal{A}^{-1}(W \odot Y)$, which is the MoHL estimate (5). The posterior mean minimizes the posterior expected squared error, hence the Bayes risk for squared error, over all measurable estimators. Each marginal posterior is a univariate Gaussian, which is symmetric about its mean, so the mean is also the posterior median and minimizes the posterior expected absolute error in every cell; this is the loss reported in our experiments. Any other estimator, including the resolvent estimator with an arbitrary free-weight operator on any support, therefore has a Bayes risk that is at least as large. □

**Corollary 2** (Further properties of a nested mixture)**.** *In the setting of Corollary 1, for every $\alpha \geq 0$: (a) the pairwise weight $L_{\mathrm{ch}}(\alpha)$ places between $v_a$ and $v_b$ is $\sum_{s \geq \max(a,b)} \alpha_s/s$; (b) for a fully observed column with uniform weight $w > 0$, the estimate $\hat{x} = (w + \mu + L_{\mathrm{ch}}(\alpha))^{-1} wx$ satisfies $\langle h_k, \hat{x} \rangle = \frac{w}{w+\mu+\Lambda_k} \langle h_k, x \rangle$ and $\langle \mathbf{1}_{e_S}, \hat{x} \rangle = \frac{w}{w+\mu} \langle \mathbf{1}_{e_S}, x \rangle$; (c) the prior $\exp(-\frac{1}{2} x^\top L_{\mathrm{ch}}(\alpha) x)$ is the sequential random-effects model $x_{v_k} \mid x_{v_1}, \ldots, x_{v_{k-1}} \sim \mathcal{N}\big(\bar{x}_{k-1}, \frac{k}{(k-1)\Lambda_k}\big)$, with $\bar{x}_{k-1}$ the mean of the $k-1$ closer sensors and a flat prior on $x_{v_1}$.*

*Proof. Part (a).* The off-diagonal entry of $P_{e_s}$ at $(v_a, v_b)$, $a \neq b$, is $-1/s$ if both vertices lie in $e_s$, that is, if $s \geq \max(a, b)$, and 0 otherwise. The weight of the pair in the adjacency $-\mathrm{offdiag}(L_{\mathrm{ch}}(\alpha))$ is therefore $\sum_{s \geq \max(a,b)} \alpha_s/s$.

*Part (b).* On the coordinates of $e_S$ the matrix $w + \mu + L_{\mathrm{ch}}(\alpha)$ has the eigenvectors of Corollary 1(i) with eigenvalues $w + \mu$ on $\mathbf{1}_{e_S}$ and $w + \mu + \Lambda_k$ on $h_k$; it is invertible because $w > 0$. Taking the inner product of $(w + \mu + L_{\mathrm{ch}}(\alpha))\hat{x} = wx$ with each eigenvector and using the symmetry of the matrix gives the two stated identities. *Part (c).* By Corollary 1(i) and the spectral decomposition of a symmetric matrix,

$$x^\top L_{\mathrm{ch}}(\alpha) x = \sum_{k=2}^{S} \Lambda_k \frac{\langle h_k, x \rangle^2}{\|h_k\|_2^2} = \sum_{k=2}^{S} \Lambda_k \, \frac{k-1}{k} \, \big(x_{v_k} - \bar{x}_{k-1}\big)^2,$$

because $\langle h_k, x \rangle = \bar{x}_{k-1} - x_{v_k}$ with $\bar{x}_{k-1} = \frac{1}{k-1} \sum_{j<k} x_{v_j}$ and $\|h_k\|_2^2 = k/(k-1)$. Hence

$$\exp\Big(-\frac{1}{2} x^\top L_{\mathrm{ch}}(\alpha) x\Big) = \prod_{k=2}^{S} \exp\Big(-\frac{(k-1)\Lambda_k}{2k} \big(x_{v_k} - \bar{x}_{k-1}\big)^2\Big).$$

For $\Lambda_k > 0$ the $k$-th factor, viewed as a function of $x_{v_k}$ with $x_{v_1}, \ldots, x_{v_{k-1}}$ fixed, is proportional to the $\mathcal{N}(\bar{x}_{k-1}, k/((k-1)\Lambda_k))$ density, and its normalizing constant does not depend on $x$. A product of conditional densities, each depending only on earlier coordinates, is the joint density of the corresponding sequential model, here with a flat (improper) density on $x_{v_1}$. If $\Lambda_k = 0$ the factor is constant and $x_{v_k}$ is unconstrained, which is the limit of infinite conditional variance.

□

**Remarks.** In the paper $W = M \odot \pi^{\odot -1}$ is a propensity weight rather than a noise precision. When the propensity is constant the two differ by a scalar, and that scalar is absorbed by the learned $(\alpha, \lambda_T, \mu)$, so Theorem 2(b) applies up to a reparameterization. Theorem 2 also explains why the hyperedge, and not free pair weights, is the natural parameterization of group-level dependence: a one-factor group of size $s$ has one dependence parameter, the ratio of factor to idiosyncratic variance, and $\alpha$ is that parameter, whereas free pair weights spend $\binom{s}{2}$ parameters to represent the same single degree of freedom.

**Learned shared weights come close to the optimum; free weights do not.** Theorem 2(b) is the case for sharing: under MoHL's Gaussian group model, MoHL at the true weights is the posterior mean, and no estimator does better, free weights on any support included. On data drawn from this model, learned shared weights come within $0.2\%$ of this oracle at every window. Free weights stay clearly above it and have the higher error in every run, although their support contains the true coupling, so their shortfall is estimation error alone; it shrinks as the window grows (Figure 14a). The model also implies that the members of a group are equally correlated. Real discovered groups come close: their within-group correlations spread far less than those of random groups of the same order, though more than exact equality allows, which Lawley's test rejects (Figure 14b). Real groups thus follow the model closely but not exactly, so the mismatch $h^2$ of Proposition 2 is small but not zero.

### D.6 Shared versus free weights

Sharing trades variance for bias: it reduces the number of parameters, but it is inaccurate when the couplings within a group differ. The comparison of Section 3.4 has an exact analogue in Gaussian-design least squares. Consider $m$ predictors, $n$ jointly observed samples, coefficient vector $\beta^\star \in \mathbb{R}^m$, and noise variance $\sigma^2$. Let $U \subset \mathbb{R}^m$ be the $r$-dimensional coefficient space imposed by sharing.

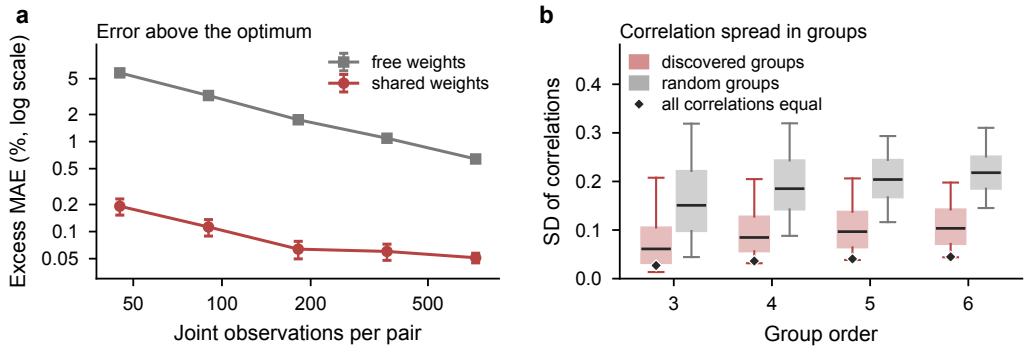


Figure 14: **The case for sharing (Theorem 2(b))**: (a) excess MAE of learned shared and free weights over the oracle on data from MoHL's model, and (b) spread of the correlations within discovered and random groups.

**Proposition 2** (Shared versus free weights)**.** *Let $x_l \sim \mathcal{N}(0, I_m)$ and $y_l = x_l^\top \beta^\star + \varepsilon_l$ with $\varepsilon_l \sim \mathcal{N}(0, \sigma^2)$, independently for $l = 1, \dots, n$; let $\hat\beta^{\text{free}}$ be least squares over $\mathbb{R}^m$ and $\hat\beta^{\text{shared}}$ least squares over $U$, and let $h^2 := \|\beta^\star - \Pi_U \beta^\star\|_2^2$ be the* group-structure mismatch*. For $n > m + 1$ the out-of-sample excess risks $\mathbb{E}\|\hat\beta - \beta^\star\|_2^2$ are*

$$\mathbb{E}\mathcal{R}(\hat\beta^{\text{free}}) = \frac{m\,\sigma^2}{n - m - 1}, \qquad \mathbb{E}\mathcal{R}(\hat\beta^{\text{shared}}) = h^2 + \frac{r\,(\sigma^2 + h^2)}{n - r - 1}, \tag{15}$$

*and sharing has the lower risk if and only if $h^2 < (m - r)\,\sigma^2/(n - m - 1)$.*

Sharing therefore helps when evidence is scarce and group mismatch is small; free weights catch up as evidence grows or the groups become inaccurate.

In the proof, $X \in \mathbb{R}^{n\times m}$ has rows $x_l^\top$, $y = X\beta^\star + \varepsilon$, and $Q \in \mathbb{R}^{m\times r}$ is an orthonormal basis of $U$, so $\Pi_U = QQ^\top$.

*Proof of Proposition 2. Step 1: a Wishart identity.* If $Z \in \mathbb{R}^{n\times k}$ has i.i.d. $\mathcal{N}(0, I_k)$ rows and $n > k + 1$, then $Z^\top Z \sim \text{Wishart}_k(n, I_k)$ and $\mathbb{E}[(Z^\top Z)^{-1}] = I_k/(n - k - 1)$ (Muirhead, 1982, Chapter 3), so $\mathbb{E}\,\text{tr}\big((Z^\top Z)^{-1}\big) = k/(n - k - 1)$.

*Step 2: the free estimator.* With $X$ of full column rank almost surely, $\hat\beta^{\text{free}} - \beta^\star = (X^\top X)^{-1} X^\top \varepsilon$. Conditioning on $X$ and using $\mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 I_n$,

$$\mathbb{E}\big[\|\hat\beta^{\text{free}} - \beta^\star\|_2^2 \;\big|\; X\big] = \sigma^2\,\text{tr}\big((X^\top X)^{-1}\big),$$

and Step 1 with $k = m$ gives $\mathbb{E}\mathcal{R}(\hat\beta^{\text{free}}) = m\sigma^2/(n - m - 1)$.

*Step 3: the shared estimator.* Least squares over $U$ is $\hat\beta^{\text{shared}} = Q\hat\gamma$ with $\hat\gamma = (Z^\top Z)^{-1} Z^\top y$ and $Z := XQ \in \mathbb{R}^{n\times r}$. Decompose $\beta^\star = Q\gamma^\star + \beta_\perp^\star$ with $\gamma^\star = Q^\top \beta^\star$ and $\beta_\perp^\star = (I - QQ^\top)\beta^\star$, $\|\beta_\perp^\star\|_2^2 = h^2$. Then $y = Z\gamma^\star + \eta$ with $\eta_l = x_l^\top \beta_\perp^\star + \varepsilon_l$. Because $x_l$ is isotropic Gaussian, $Q^\top x_l$ and $(I - QQ^\top)x_l$ are independent, so $\eta$ is independent of $Z$ with i.i.d. $\mathcal{N}(0, \sigma^2 + h^2)$ entries: the part of $\beta^\star$ that $U$ cannot express acts as additional noise. Hence, as in Step 2, $\mathbb{E}\|\hat\gamma - \gamma^\star\|_2^2 = (\sigma^2 + h^2)\,r/(n - r - 1)$ by Step 1 with $k = r$. Finally $\hat\beta^{\text{shared}} - \beta^\star = Q(\hat\gamma - \gamma^\star) - \beta_\perp^\star$ with the two terms orthogonal, so

$$\mathbb{E}\mathcal{R}(\hat\beta^{\text{shared}}) = \mathbb{E}\|\hat\gamma - \gamma^\star\|_2^2 + h^2 = h^2 + \frac{r(\sigma^2 + h^2)}{n - r - 1}.$$

*Step 4: comparison.* Sharing wins when $h^2\big(1 + \frac{r}{n-r-1}\big) < \sigma^2\big(\frac{m}{n-m-1} - \frac{r}{n-r-1}\big)$. The left factor is $h^2\,\frac{n-1}{n-r-1}$ and the right bracket equals $\frac{(m-r)(n-1)}{(n-m-1)(n-r-1)}$, so after cancelling $\frac{n-1}{n-r-1} > 0$ the condition is exactly $h^2 < (m - r)\sigma^2/(n - m - 1)$. □

**Corollary 3** (Scarcity crossover)**.** *Sharing has lower expected risk than the free model if and only if*

$$n < n^\star := m + 1 + \frac{(m-r)\,\sigma^2}{h^2}, \tag{16}$$

*with $n^\star = \infty$ when $h^2 = 0$. Under an independent Bernoulli observation model with rate $\pi$, in which a pair is jointly observed $n \asymp T\pi^2$ times, the crossover occurs at a window length $T^\star \asymp \big(m+1+(m-r)\sigma^2/h^2\big)/\pi^2$: it grows with the number of couplings, with the noise, and with the observation deficit, and shrinks with the mismatch.*

*Proof.* For $n > m+1$ the condition of Step 4 is $h^2(n-m-1) < (m-r)\sigma^2$, i.e. $n < m+1+(m-r)\sigma^2/h^2$ when $h^2 > 0$, and it holds for every $n$ when $h^2 = 0$. For $n \le m+1$ the free least-squares problem has no finite-risk solution while the shared one does (it needs only $n > r+1$), so sharing wins there as well. For independent Bernoulli masks the number of timesteps at which a fixed pair is jointly observed is $\mathrm{Binomial}(T, \pi^2)$ and concentrates around $T\pi^2$ (multiplicative Chernoff bound), which gives $T^\star$. □

**Reading the corollary.** Equation (15) is the bias–variance decomposition of a restricted least-squares estimator under a random design: the restriction to $U$ costs an approximation error $h^2$ that does not shrink with $n$, and it replaces a variance of order $m\sigma^2/(n-m)$ by one of order $r\sigma^2/(n-r)$. Three monotonicities follow from (16). The crossover $n^\star$ grows linearly with the number of couplings $m$ that one order weight replaces: the larger the pairs governed by one mixture weight, the longer sharing remains preferable. It grows with the noise level $\sigma^2$. And it shrinks with the mismatch $h^2$: if the true couplings within an interaction order are nearly equal, sharing is nearly costless and wins at every realistic sample size. In the comparison of Section 3.4, $m$ is the number of pairwise couplings on the discovered support, $r = S-1$ is the number of order weights, $\beta^\star$ plays the role of the ideal pair weights on that support, and $n$ is governed by the window length and the observation rate; Table 9 varies $n$ through the window.

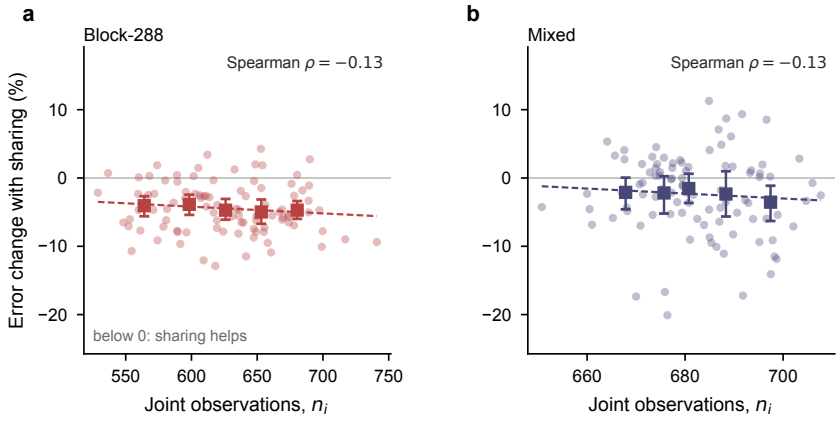


Figure 15: **Sharing and joint evidence (Proposition 2)**: each sensor's change in MAE from free to shared weights against its joint evidence $n_i$, under (a) Block-288 and (b) Mixed.

**Within one network, the evidence trend is not detectable.** Proposition 2 predicts that sharing helps most where the evidence per coupling is scarcest. Comparing windows changes the data as well as the evidence, whereas within a run sensors differ in joint evidence while data, mask and training stay fixed. Across the sensors of the five Block-288 and five Mixed runs of Figure 6, sharing lowers the error of the average sensor, but the size of its gain is not associated with the sensor's joint evidence $n_i$, the median number of joint observations with its discovered neighbors: the rank correlation is weak, and its sign varies across the five seeds (Figure 15). The test has little power, because $n_i$ varies little within one network. The predicted trend is thus borne out on data drawn from the model (Appendix D.5) but not detectable across the sensors of one network.

### D.7 THE ONE-ADJOINT GRADIENT

Section 3.3 uses the adjoint gradients (8) for every parameter. We record the statement and its proof once, in the generality that also covers the free pair weights of the free model and the gate parameters of Section 3.3.

**Proposition 3** (One-adjoint gradient)**.** *Let $\widehat{X}(\theta)$ solve $\mathcal{A}(\theta)\widehat{X} = b$, where $\mathcal{A}(\theta)$ is symmetric positive definite and continuously differentiable in $\theta$, and let $\mathcal{L}(\widehat{X})$ be any differentiable outer loss. Define the adjoint state $\Psi$ by $\mathcal{A}(\theta)\Psi = \partial\mathcal{L}/\partial\widehat{X}$. Then for any scalar parameter $\theta_r$ that enters $\mathcal{A}$ but not $b$,*

$$\frac{\partial \mathcal{L}}{\partial \theta_r} = -\Big\langle \Psi, \frac{\partial \mathcal{A}}{\partial \theta_r}\widehat{X} \Big\rangle_F. \tag{17}$$

*Proof. Step 1: differentiability of the solution map.* Define $F(\theta, X) = \mathcal{A}(\theta)X - b$. The map $F$ is continuously differentiable, and its partial derivative with respect to $X$ is the linear map $\mathcal{A}(\theta)$, which is invertible because it is symmetric positive definite. By the implicit function theorem the solution $\widehat{X}(\theta) = \mathcal{A}(\theta)^{-1}b$ is continuously differentiable, and differentiating $\mathcal{A}(\theta)\widehat{X}(\theta) = b$ with respect to a scalar parameter $\theta_r$ that does not enter $b$ gives

$$\frac{\partial \mathcal{A}}{\partial \theta_r}\widehat{X} + \mathcal{A}\,\frac{\partial \widehat{X}}{\partial \theta_r} = 0, \qquad \text{hence} \qquad \frac{\partial \widehat{X}}{\partial \theta_r} = -\mathcal{A}^{-1}\Big(\frac{\partial \mathcal{A}}{\partial \theta_r}\widehat{X}\Big).$$

*Step 2: chain rule.* With $G = \partial\mathcal{L}/\partial\widehat{X}$, $\partial\mathcal{L}/\partial\theta_r = \langle G, \partial\widehat{X}/\partial\theta_r\rangle_F = -\langle G, \mathcal{A}^{-1}(\partial_{\theta_r}\mathcal{A}\,\widehat{X})\rangle_F$.

*Step 3: moving the inverse.* The inverse of a symmetric invertible map is symmetric: for any $U, V$, setting $U = \mathcal{A}U'$ and $V = \mathcal{A}V'$ gives $\langle U, \mathcal{A}^{-1}V\rangle_F = \langle \mathcal{A}U', V'\rangle_F = \langle U', \mathcal{A}V'\rangle_F = \langle \mathcal{A}^{-1}U, V\rangle_F$. With $\Psi = \mathcal{A}^{-1}G$ this turns Step 2 into (17). The adjoint state $\Psi$ does not depend on $r$, so a single solve serves every parameter. □

**Instantiation for every parameter used in the paper.** The operator is $\mathcal{A}Z = W \odot Z + \sum_k \alpha_k \bar{L}_k Z + \lambda_T Z L_T + \mu Z$ with $\bar{L}_0 = L_{\text{prior}}$, $\bar{L}_s = L_{\mathcal{H}_s}$ and $\alpha_k = e^{a_k}$. Differentiating, $\partial_{a_k}\mathcal{A}\,Z = \alpha_k \bar{L}_k Z$, $\partial_{\log\lambda_T}\mathcal{A}\,Z = \lambda_T Z L_T$ and $\partial_{\log\mu}\mathcal{A}\,Z = \mu Z$, and substituting into (17) gives (8); the factors $\alpha_k$, $\lambda_T$, $\mu$ are the chain-rule factors of the logarithmic parameterization, which keeps every parameter positive without constraints. For the free model (10), a single pair weight $(A_{\text{free}})_{ij}$ enters the spatial operator through $\alpha_1(A_{\text{free}})_{ij}L_{ij}$, so $\partial\mathcal{L}/\partial(A_{\text{free}})_{ij} = -\alpha_1\langle\Psi, L_{ij}\widehat{X}\rangle_F = -\alpha_1(m_{ii} + m_{jj} - m_{ij} - m_{ji})$ with $m = \Psi\widehat{X}^\top$, followed by the derivative of the softplus reparameterization; all pair gradients are read off one $N \times N$ matrix. The gate parameters enter $\mathcal{A}$ through the time-varying edge weights (9) and are differentiated in the same way, with the same $\Psi$ (Appendix C.4).

**The right-hand side and the cost.** For the objective (7), $G$ has entries $\ell'_\delta(\widehat{X}_{it} - \tilde{Y}_{it} + P_{it})/|\Omega_{\text{learn}}|$ on $\Omega_{\text{learn}}$ and zero elsewhere, with $\ell'_\delta(r) = \max(-\delta, \min(\delta, r))$; the $\ell_1$ term contributes $\gamma\alpha_s$ to $\partial/\partial a_s$ directly. Since $\Omega_{\text{learn}}$ is disjoint from the fitting mask, $G$ is supported on cells where $W$ vanishes, and the adjoint system must be solved with $G$ itself as its right-hand side, not with $W \odot G$. One gradient evaluation requires the forward solve, the adjoint solve with the same operator (hence the same conditioning, Lemma 4), and one inner product $\langle\Psi, \bar{L}_k\widehat{X}\rangle_F$ per component; differentiating the solve separately for each of $p$ parameters would cost $p$ solves. For a nonsymmetric system the same argument requires the transposed operator; here $\mathcal{A} = \mathcal{A}^\top$, so the forward and adjoint solves share the operator, its preconditioner and its implementation.

### D.8 GENERALIZATION OF THE LEARNED OPERATOR

**Theorem 3** (Generalization of the learned operator)**.** *Condition on the masks and on all data in $\Omega_{\text{fit}}$, and suppose the noise at the $n = |\Omega_{\text{learn}}|$ held-out cells is independent, mean-zero and $\sigma$-sub-Gaussian. Let $\{\mathcal{A}(\theta) : \theta \in \Theta\}$ be an operator class on a box $\Theta \subset \mathbb{R}^p$ with $\mathcal{A}(\theta) \succeq \mu_{\min} I$ and $\|\mathcal{A}(\theta) - \mathcal{A}(\theta')\|_{\text{op}} \le G\|\theta - \theta'\|_1$, let $R(\theta)$ be the expected held-out Huber loss of $\hat{X}(\theta)$, and let $\hat{\theta}$ minimize* (7)*, with any nonnegative penalty* pen*, up to an optimization error $\varepsilon_{\text{opt}}$. Then with*

*probability at least* $1-\eta$*,*

$$R(\hat{\theta}) \le \inf_{\theta\in\Theta}\big\{R(\theta)+\mathrm{pen}(\theta)\big\} + C\,\delta\sigma\sqrt{\frac{p\log(1+\kappa_\Theta)+\log(2/\eta)}{n}} + \varepsilon_{\mathrm{opt}}, \tag{18}$$

*where* $C$ *is an absolute constant and* $\kappa_\Theta := p\,d_{\max}G\,\|W\odot(\tilde{Y}-P)\|_F/(\mu_{\min}^2\sigma)$*, with* $d_{\max}$ *the largest side of* $\Theta$*, enters only through the logarithm.*

We first state the stability bound on which the proof rests, then the setting precisely, then the proof, and finally the corollary that compares the shared and free guarantees.

**Lemma 6** (Resolvent stability)**.** *Let* $\mathcal{A}, \mathcal{A}'$ *be symmetric linear maps on* $\mathbb{R}^{N\times T}$ *with* $\mathcal{A}\succeq\mu I$ *and* $\mathcal{A}'\succeq\mu' I$*,* $\mu,\mu'>0$*, and let* $b\in\mathbb{R}^{N\times T}$*. Then*

$$\big\|\mathcal{A}^{-1}b-\mathcal{A}'^{-1}b\big\|_F \le \frac{\|\mathcal{A}-\mathcal{A}'\|_{\mathrm{op}}}{\mu\mu'}\,\|b\|_F.$$

*Proof.* The resolvent identity $\mathcal{A}^{-1}-\mathcal{A}'^{-1}=\mathcal{A}^{-1}(\mathcal{A}'-\mathcal{A})\mathcal{A}'^{-1}$ follows by multiplying out the right-hand side. A symmetric map with $\mathcal{A}\succeq\mu I$ has all eigenvalues at least $\mu$, so $\|\mathcal{A}^{-1}\|_{\mathrm{op}}\le 1/\mu$, and likewise $\|\mathcal{A}'^{-1}\|_{\mathrm{op}}\le 1/\mu'$. Submultiplicativity of the operator norm gives the claim. □

**Setting.** Fix the masks, and hence the disjoint sets $\Omega_{\mathrm{fit}}$ and $\Omega_{\mathrm{learn}}$, and condition on all data observed on $\Omega_{\mathrm{fit}}$. The standardization, the daily profile $P$, the affinities, the groups and the components $L_{\mathcal{H}_s}$ are computed from the fitting mask only (Section 3.3), so under this conditioning they are deterministic, and for every $\theta$ the estimate $\widehat{X}(\theta)=\mathcal{A}(\theta)^{-1}b$ with $b=W\odot(\tilde{Y}-P)$ is a deterministic matrix. On the held-out cells the target is $Z_{it}:=\tilde{Y}_{it}-P_{it}=X^\circ_{it}+\xi_{it}$, where $X^\circ$ is the standardized, profile-removed truth and the $\xi_{it}$, $(i,t)\in\Omega_{\mathrm{learn}}$, are independent, mean-zero, independent of the fitting data, and sub-Gaussian with $\|\xi_{it}\|_{\psi_2}\le\sigma$. The empirical and expected held-out risks are

$$\widehat{R}(\theta)=\frac{1}{n}\sum_{(i,t)\in\Omega_{\mathrm{learn}}}\ell_\delta\big(\widehat{X}_{it}(\theta)-Z_{it}\big),\qquad R(\theta)=\mathbb{E}_\xi\widehat{R}(\theta),$$

and $\hat{\theta}$ satisfies $\widehat{R}(\hat{\theta})+\mathrm{pen}(\hat{\theta})\le\inf_\Theta\{\widehat{R}+\mathrm{pen}\}+\varepsilon_{\mathrm{opt}}$. The box is $\Theta=\prod_{r=1}^p[\theta_r^{\mathrm{lo}},\theta_r^{\mathrm{hi}}]$ with largest side $d_{\max}$.

*Proof of Theorem 3. Step 1: from the supremum to the oracle inequality.* Let $\Delta_n=\sup_{\theta\in\Theta}|\widehat{R}(\theta)-R(\theta)|$. For any $\theta^\circ\in\Theta$,

$$\begin{aligned}R(\hat{\theta}) &\le R(\hat{\theta})+\mathrm{pen}(\hat{\theta}) \le \widehat{R}(\hat{\theta})+\mathrm{pen}(\hat{\theta})+\Delta_n\\ &\le \widehat{R}(\theta^\circ)+\mathrm{pen}(\theta^\circ)+\varepsilon_{\mathrm{opt}}+\Delta_n \le R(\theta^\circ)+\mathrm{pen}(\theta^\circ)+\varepsilon_{\mathrm{opt}}+2\Delta_n,\end{aligned}$$

using $\mathrm{pen}\ge 0$, the definition of $\Delta_n$, the near-optimality of $\hat{\theta}$, and the definition of $\Delta_n$ again. Taking the infimum over $\theta^\circ$ reduces the theorem to a bound on $\Delta_n$.

*Step 2: pointwise concentration.* Fix $\theta$ and write $u_{it}=\widehat{X}_{it}(\theta)-X^\circ_{it}$, so that the summands of $\widehat{R}(\theta)$ are $f_{it}(\xi_{it})=\ell_\delta(u_{it}-\xi_{it})$. The Huber loss is $\delta$-Lipschitz, so $V_{it}=f_{it}(\xi_{it})-f_{it}(0)$ satisfies $|V_{it}|\le\delta|\xi_{it}|$ and therefore $\|V_{it}\|_{\psi_2}\le\delta\sigma$. Centering costs an absolute constant, $\|V_{it}-\mathbb{E}V_{it}\|_{\psi_2}\le C_0\delta\sigma$, and $f_{it}(\xi_{it})-\mathbb{E}f_{it}(\xi_{it})=V_{it}-\mathbb{E}V_{it}$. The summands are independent, so the general Hoeffding inequality for sub-Gaussian variables gives, for an absolute constant $c>0$ and every $t>0$,

$$\Pr\big[|\widehat{R}(\theta)-R(\theta)|\ge t\big]\le 2\exp\Big(-\frac{c\,n\,t^2}{\delta^2\sigma^2}\Big).$$

*Step 3: Lipschitz dependence on* $\theta$*.* By Lemma 6 with $\mu=\mu'=\mu_{\min}$ and the hypothesis on the class,

$$\|\widehat{X}(\theta)-\widehat{X}(\theta')\|_F\le\frac{G\|b\|_F}{\mu_{\min}^2}\,\|\theta-\theta'\|_1=:L_\Theta\|\theta-\theta'\|_1.$$

Since $\ell_\delta$ is $\delta$-Lipschitz and $\sum_{\Omega_{\mathrm{learn}}}|v_{it}|\le\sqrt{n}\,\|v\|_F$,

$$|\widehat{R}(\theta)-\widehat{R}(\theta')|\le\frac{\delta}{n}\sum_{(i,t)\in\Omega_{\mathrm{learn}}}|\widehat{X}_{it}(\theta)-\widehat{X}_{it}(\theta')|\le\frac{\delta L_\Theta}{\sqrt{n}}\,\|\theta-\theta'\|_1,$$

and the same bound holds for $R = \mathbb{E}_\xi \widehat{R}$.

*Step 4: covering.* Cover each side of $\Theta$ by points spaced $2\varepsilon/p$ apart. The resulting grid $\mathcal{N}_\varepsilon$ has at most $(1 + p\, d_{\max}/(2\varepsilon))^p$ points, and every $\theta \in \Theta$ lies within $\varepsilon/p$ of a grid point in each coordinate, hence within $\varepsilon$ in the $\ell_1$ norm. By Step 3,

$$\Delta_n \le \max_{\theta \in \mathcal{N}_\varepsilon} |\widehat{R}(\theta) - R(\theta)| + \frac{2\delta L_\Theta \varepsilon}{\sqrt{n}}.$$

Choose $\varepsilon = \sigma/(2L_\Theta)$, so that the second term is $\delta\sigma/\sqrt{n}$ and $\log|\mathcal{N}_\varepsilon| \le p\log(1 + p\, d_{\max} L_\Theta/\sigma) = p\log(1+\kappa_\Theta)$ with $\kappa_\Theta = p\, d_{\max} G\|b\|_F/(\mu_{\min}^2\sigma)$.

*Step 5: union bound.* Applying Step 2 to every point of $\mathcal{N}_\varepsilon$ with $t = \delta\sigma\sqrt{(\log|\mathcal{N}_\varepsilon| + \log(2/\eta))/(cn)}$ and taking a union bound, with probability at least $1-\eta$,

$$\Delta_n \le \delta\sigma\sqrt{\frac{p\log(1+\kappa_\Theta) + \log(2/\eta)}{c\,n}} + \frac{\delta\sigma}{\sqrt{n}} \le \frac{C}{2}\,\delta\sigma\sqrt{\frac{p\log(1+\kappa_\Theta) + \log(2/\eta)}{n}}$$

for an absolute constant $C$, where the last step uses $\log(2/\eta) \ge \log 2$. Combining with Step 1 gives (18). □

**Lemma 7** (Both model classes satisfy the hypotheses)**.** *On any box on which $a_k \le \bar{a}$, $\log\lambda_T \le \bar{l}_T$ and $\log\mu \in [\log\mu_{\min}, \bar{l}_\mu]$, the MoHL class $\theta = (a_0, a_2, \dots, a_S, \log\lambda_T, \log\mu)$ satisfies the hypotheses of Theorem 3 with $p = S+2$ and $G = \max(e^{\bar{a}}, 4e^{\bar{l}_T}, e^{\bar{l}_\mu})$. On a box that in addition bounds the pair logits, $|\vartheta_{ij}| \le \bar{\vartheta}$, the free class $\theta = (a_0, a_1, (\vartheta_{ij})_{\{i,j\}\in E}, \log\lambda_T, \log\mu)$ satisfies them with $p = m+4$ and a finite $G$ depending on $\bar{a}$, $\bar{\vartheta}$ and the row sums of $A_{\text{shared}}$.*

*Proof.* In both classes $\mathcal{A}(\theta) \succeq \mu I \succeq \mu_{\min} I$ by Lemma 4. For MoHL, writing $\bar{L}_0 := L_{\text{prior}}$ and $\bar{L}_s := L_{\mathcal{H}_s}$,

$$\mathcal{A}(\theta)Z - \mathcal{A}(\theta')Z = \sum_k (\alpha_k - \alpha_k')\bar{L}_k Z + (\lambda_T - \lambda_T')ZL_T + (\mu - \mu')Z.$$

The maps $Z \mapsto \bar{L}_k Z$, $Z \mapsto ZL_T$ and the identity have operator norms $\rho(\bar{L}_k) = 1$, $\rho(L_T) < 4$ and 1, and the mean value theorem gives $|e^x - e^{x'}| \le e^{\bar{x}}|x - x'|$ for $x, x' \le \bar{x}$. Summing the three contributions yields the stated $G$. For the free class the spatial part is $\alpha_0\bar{L}_0 + \alpha_1\mathcal{L}(A_{\text{free}}(\vartheta))$ with $\mathcal{L}(B) = D_B - B$. Gershgorin's theorem gives $\|\mathcal{L}(B)\|_{\text{op}} \le 2\max_i \sum_j |B_{ij}|$, the softplus is 1-Lipschitz, and $A_{\text{free}}(\vartheta) - A_{\text{free}}(\vartheta') = A_{\text{shared}} \odot (\text{softplus}(\vartheta) - \text{softplus}(\vartheta'))/\,\text{softplus}(0)$. Hence $\|\mathcal{L}(A_{\text{free}}(\vartheta)) - \mathcal{L}(A_{\text{free}}(\vartheta'))\|_{\text{op}} \le (2\max_{ij}(A_{\text{shared}})_{ij}/\,\text{softplus}(0))\|\vartheta - \vartheta'\|_1$, while $\|\mathcal{L}(A_{\text{free}}(\vartheta))\|_{\text{op}}$ is bounded on the box by $2\,\text{softplus}(\bar{\vartheta})\max_i \sum_j (A_{\text{shared}})_{ij}/\,\text{softplus}(0)$. The product rule applied to $\alpha_1\mathcal{L}(A_{\text{free}}(\vartheta))$ then gives a finite $G$. □

**Corollary 4** (Comparison of the shared and free guarantees)**.** *Let $R^\star_{\text{shared}}$ and $R^\star_{\text{free}}$ be the infima of $R$ over the two classes on the same relational support, and let $\Delta = R^\star_{\text{shared}} - R^\star_{\text{free}}$. If the free box is large enough to contain the ratios $c_{ij}$ of Proposition 1, then $\Delta \ge 0$. Writing $\epsilon_n(p,\kappa) = C\delta\sigma\sqrt{(p\log(1+\kappa) + \log(2/\eta))/n}$ for the estimation term of* (18) *and neglecting the penalties, the bound of Theorem 3 for the shared class is smaller than the bound for the free class if and only if*

$$\Delta < \epsilon_n(m+4, \kappa_{\text{free}}) - \epsilon_n(S+2, \kappa_{\text{shared}}).$$

*For $m \gg S$ and $\Delta > 0$ this holds for all $n \le n^\star$ with $n^\star \asymp \delta^2\sigma^2\, m\log(1+\kappa_{\text{free}})/\Delta^2$, and for every $n$ when $\Delta = 0$.*

*Proof.* By Proposition 1 every shared operator is a limit of free operators on the same support, and by Lemma 6 the risk $R$ is continuous in the operator, so the free infimum cannot exceed the shared one; this gives $\Delta \ge 0$. The two bounds are $R^\star_{\text{shared}} + \epsilon_n(S+2, \kappa_{\text{shared}})$ and $R^\star_{\text{free}} + \epsilon_n(m+4, \kappa_{\text{free}})$, and subtracting gives the stated condition. Both estimation terms scale as $n^{-1/2}$, so the right-hand side equals $c_*/\sqrt{n}$ with $c_* \asymp \delta\sigma\sqrt{m\log(1+\kappa_{\text{free}})}$ when $m \gg S$, and $\Delta < c_*/\sqrt{n}$ is equivalent to $n < c_*^2/\Delta^2$. □

## E LIMITATIONS AND FUTURE WORK

**Structure learned from the data at hand.** MoHL learns its groups and their weights from the same incomplete window it imputes, so it needs no external training data and adapts to each deployment. A natural extension is to combine this observed evidence with relational knowledge from beyond the window, such as longer histories, related deployments or network topology, and to learn how much weight each source deserves.

**Evolving groups.** Within a window, the gate already adapts the learned structure to local missingness. Allowing the groups themselves to form, dissolve or reweight over time, driven by the same pairwise evidence, would extend the mixture to relations that shift with incidents, weather or the time of day, while keeping its parameter count small.

**Sharing as a continuum.** Our results show that structured sharing reduces error across regimes and window lengths. Shared and free weights can be seen as the two ends of a continuum: hierarchical priors that shrink each pair weight toward its group weight according to the pair's joint evidence would let a model choose its place on it automatically. The same idea applies wherever relations are learned from unevenly observed data.

**Complementing deep imputers.** Our protocol trains every method on the same window, which isolates what each learns from the data at hand. The structure MoHL learns is compact and interpretable, and it can serve deep imputers directly: the mixture Laplacian could act as their graph or as a structural prior, combining evidence-efficient relational structure with rich nonlinear dynamics.

**Beyond traffic networks.** Discovering groups from pairwise evidence and sharing one weight per group order is not specific to traffic. The same principle applies to other networked measurements with structured missingness, such as air quality, energy and environmental monitoring, and extends to larger networks through sparse group construction and preconditioned solvers. A theory of relational evidence under spatially and temporally dependent missingness would further sharpen when higher-order sharing pays.