EXACT GROUP RECOVERY FROM SUBQUADRATIC OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering a finite group operation table from partial observations presents a fundamentally unusual completion task. Unlike classical matrix completion, which relies on a low-rank prior and geometric incoherence to interpolate missing entries, completing a group table requires enforcing an associativity prior and exploiting the discrete Latin-square property, which guarantees uniform marginals without structural incoherence assumptions. We resolve this by analyzing a canonical operator-valued factorization model under gradient-sensitivity regularization. This objective forces the model's latent matrix embeddings to become unitary and satisfy the exact multiplication laws of the observed pairs once a basic coverage threshold of $n \log(3n/\delta)$ observations is reached (coupon collector). By identifying the structural observation patterns required to propagate these localized relations into exact global recovery, we establish sub-quadratic sample complexity bounds under two regimes. Under active query selection, a presentation-dependent scaffold deterministically recovers the complete table in at most $3n - 2 + |R_{\text{indep}}|$ queries, where $|R_{\text{indep}}|$ counts the pairs needed to close the relator cycles, by tracing a Cayley spanning tree and the group's defining relators to algebraically synchronize the representations. Under passive Bernoulli sampling, $\mathcal{O}(n^{3/2}\sqrt{\log n})$ expected observations guarantee exact table recovery with high probability, as overlapping observations form connected graphs that globally propagate the associativity prior. In both regimes, evaluating all negative candidate outputs for just one observed input pair provides the necessary anchor to eliminate degenerate solutions and select the regular representation. These bounds establish the first closed theory of partial-observation learning for general finite groups, verified formally in Lean 4 and empirically across 32 group families.

1 INTRODUCTION AND MOTIVATION

Learning finite operation tables provides a concrete setting for studying whether observed examples determine an underlying algebraic rule. For a finite group G of order n, the task is to recover all n^2 products from a subset of labeled input pairs. Group laws relate observed and unobserved products, but agreement with the observations alone does not establish that a learned model obeys those laws. A recovery theory must explain both how the reconstruction objective enforces structure and which observation patterns make that structure identify the target.

Restricted isometry and related conditions provide this connection in sparse and low-rank recovery: a condition on the measurement operator guarantees recovery by a specified reconstruction program, and a sample complexity theorem bounds the number of measurements sufficient for that condition to hold. Examples include ℓ_1 and nuclear-norm minimization, as well as recovery under explicit sparsity or rank constraints (Recht et al., 2010). Restricted approximate invertibility provides a related connection between one-bit measurements and iterative reconstruction (Friedlander et al., 2022). We develop an algebraic counterpart for finite multiplication tables, connecting a regularized factorization objective, sufficient conditions on the observations, and the number of observations needed for exact recovery.

Our reconstruction model is HyperCube, an operator-valued tensor factorization with gradient-sensitivity regularization. Huh et al. (2025) characterize its global minimizers for fully observed tables. We study partial observations. Under coverage of every symbol in each of the two input

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.

slots and the output slot, the objective's equality conditions force every exact global minimizer, after a common-scale normalization, to have unitary slices satisfying the observed multiplication laws. Recovery then requires sufficient relations among the observations to determine the unobserved products. We establish two such conditions: an actively selected algebraic scaffold and connectivity of relative-comparison graphs formed by random observations.

This paper addresses three central theoretical questions: (1) identifying the structural conditions on the observation set that guarantee unique Cayley table recovery; (2) bounding the sample complexity required to prevent degenerate parameter collapse and secure full rank; and (3) characterizing why group recovery requires only sub-quadratic observations.

In this work, we resolve these questions by establishing the first closed theory of partial-observation learning and sufficient subquadratic sample complexity bounds for the HyperCube model. Our presentation-dependent active query bound and passive Bernoulli recovery bound (Section 1.1) are the first closed-form structural sample complexity upper bounds that apply to all finite groups (and their isotopes), including non-commutative groups such as symmetric groups and dihedral groups. Recovery thresholds measured on 32 groups are consistent with growth of order $\tilde{\mathcal{O}}(n^{3/2})$, the order of the passive bound (Section 5).

1.1 Summary of results

To overcome the $\Omega(n^2)$ memorization baseline required to specify an arbitrary quasigroup, we establish a hierarchy of sub-quadratic sample complexity bounds for exact group recovery:

- **Metric Locking (Tier 1):** Under passive uniform sampling, $m \geq n \log(3n/\delta)$ observations reach the three-slot coupon-collector threshold with probability at least $1 - \delta$, which forces every exact global minimizer into full-rank, scaled-unitary slices (Theorems 3, 4 and A.7).
- **Active Presentation Scaffolds:** Under active query selection, observing an identity cross, a Cayley spanning tree, and the pairs tracing the presentation relators ($S_{\min}$) deterministically recovers the full Cayley table in at most $\mathbf{3n - 2 + |R_{\text{indep}}|}$ queries (Theorem 6), where $|R_{\text{indep}}|$ counts the pairs needed to close the relator cycles of a chosen presentation; a single complete output fiber selects the regular representation (Theorem 7).
- **Passive Bernoulli Recovery:** For a fixed group or group isotope of order n, under passive Bernoulli sampling, $\mathbf{\sqrt{6}\, n^{3/2} (\log(8n^2/\delta))^{1/2}}$ expected observed pairs suffice, with probability at least $1 - \delta$, for every exact global minimizer of the regularized HyperCube objective (5), with $\lambda > 0$, to recover the entire labeled table, including all unobserved products (Corollaries 10 and 11 and Appendix D).

Formal verification. For finite groups in the complex-matrix model, a Lean 4 development formalizes metric locking, character recovery, presentation-scaffold recovery, and passive Bernoulli recovery; Appendix M states its exact scope and the results that lie outside it.[1]

2 Related work

Algorithmic Operation Discovery. Power et al. (2022) introduced the binary operation table recovery task to machine learning and first reported the grokking phenomenon. Subsequent studies (Nanda et al., 2023; Gromov, 2023; Varma et al., 2023; Zhong et al., 2023) demonstrated that generalization in these tasks involves discovering irreducible unitary group characters. Huh (2025) showed that such generalization involves discovering exact unitary group representations in operator-valued factorization networks. Recently, Mohamadi et al. (2024) bounded the sample complexity of two-layer quadratic networks on modular addition by $\widetilde{\mathcal{O}}(p^{5/3})$; their analysis relies on discrete Fourier features restricted to Abelian cyclic groups ($\mathbb{Z}_p$). Our work provides the first a-priori sample complexity bounds for general finite groups: gradient sensitivity regularization forces factor slices into unitary group representations via the Matrix AM–GM inequality.

Low-Rank Matrix Completion. Matrix completion guarantees combine a low-rank prior with incoherence conditions (Candès & Recht, 2009; Ge et al., 2016). Group table completion replaces

[1] https://anonymous.4open.science/r/HyperCube_Group_LEAN-8E63/

them by an associativity prior and the Latin-square property of quasigroups (van Lint & Wilson, 2001, Ch. 17), which yields three-slot coverage without incoherence assumptions (Theorem A.7); Appendix N gives the full comparison.

Teaching Dimension and Property Testing. Teaching Dimension is the minimum number of examples an oracle must provide to uniquely specify a target concept (Goldman & Kearns, 1995). Algebraic property testing has analyzed randomized linearity testers (Blum et al., 1993) and spot-checkers for group and field operations (Ergün et al., 2000). We move from probabilistic testing to exact algebraic recovery with presentation-dependent teaching sets, whose size is governed directly by the group's generators and presentation relators.

3 Problem Setting

3.1 Completion task

Consider a relational structure $a \circ b = c$, where $a, b, c \in Q$ are symbols from a finite vocabulary of order $|Q| = n$, and $\circ : Q \times Q \to Q$ is an unknown binary operation (Power et al., 2022). An algebraic completion task recovers the full operation table $(Q, \circ)$ from observed positive triples (a, b, c) with $a \circ b = c$; the table is the indicator 3-tensor $\delta \in \{0,1\}^{n\times n\times n}$, $\delta_{abc} = \mathbf{1}_{\{a\circ b=c\}}$.

Let $S \subseteq Q \times Q$ be a set of observed input pairs. It defines the observed positive support $\Omega_S^+ := \{(a, b, a \circ b) : (a, b) \in S\}$, with target value 1, and the full observed support $\Omega_S := \{(a, b, c) : (a, b) \in S,\ c \in Q\}$ across all output candidates. We count observations as queried input pairs; the n candidate values of an observed pair are not counted separately. We write m for the number of independent draws with replacement, $|S| \le m$ for the number of distinct observed pairs, $\bar{m} = pn^2$ for the expected size of a Bernoulli-p mask, and m^* for a measured threshold.

3.2 Operator-valued factorization and gradient sensitivity objective

We analyze the operator-valued factorization model of Huh (2025), $T : Q^3 \to \mathbb{C}$ defined as a trilinear product of three families of complex matrices $\Theta = (A, B, C) \in (\mathbb{C}^{n\times n})^{3n}$ indexed by $a, b, c \in Q$:

$$T_{abc}(\Theta) = \frac{1}{n} \operatorname{Tr}(A_a B_b C_c). \tag{1}$$

Define the per-triple scalar sensitivity

$$r_\Theta(a, b, c) := n\|\nabla_\Theta T_{abc}(\Theta)\|_F^2 = \frac{1}{n}\left(\|A_a B_b\|_F^2 + \|B_b C_c\|_F^2 + \|C_c A_a\|_F^2\right) \tag{2}$$

Note that $r_\Theta = 3$ at any unitary triple. The sampled regularizer is

$$\mathscr{H}_S(\Theta) := \sum_{(a,b)\in S} r_\Theta(a, b, a \circ b). \tag{3}$$

The minimum of $\mathscr{H}_S$ is $3|S|$. The *positive program*

$$(\mathrm{P}_+) \qquad \min_\Theta \mathscr{H}_S(\Theta) \quad \text{subject to} \quad T_{ab,a\circ b}(\Theta) = 1 \quad ((a, b) \in S) \tag{4}$$

imposes only the observed positive triples. Choosing an observed pair $(a_0, b_0) \in S$, possibly after seeing S, and imposing its *full-output constraints* $T_{a_0 b_0 c}(\Theta) = \mathbf{1}_{\{a_0\circ b_0=c\}}$ for all $c \in Q$ gives an anchored program $(\mathrm{P}_{\mathrm{anc}})$; such a pair is an *anchor*, which specifies one output vector at one input pair, not a row or column of the table. Full-output constraints at every observed pair give the *exact-fit program*, $T_{abc}(\Theta) = \delta_{abc}$ on Ω_S. Finally, for $\lambda > 0$ the finite-regularization objective is

$$\mathcal{J}_{\lambda,\Omega_S}(\Theta) := \sum_{(a,b)\in S}\Big[\sum_{c\in Q}\big|\delta_{abc} - T_{abc}(\Theta)\big|^2 + \lambda r_\Theta(a, b, a \circ b)\Big], \tag{5}$$

the sum of the positive fitting error, the sensitivity, and the negative-target error $\sum_{c\neq a\circ b} |T_{abc}|^2$. The positive fitting term coupled with sensitivity drives **metric locking and algebraic homomorphism** via the Matrix AM–GM inequality, while the negative-target terms enforce **character orthogonality**

and faithfulness, precluding degenerate states such as the trivial representation $A_g = s_\lambda I_n$ (Proposition A.4). The experiments (Section 5) minimize (5) with sum reduction and this normalization of r_Θ; implementation-specific normalizations are in Appendix K.

Our analysis separates two questions: which local matrix relations the objective enforces on observed examples, and which observation patterns extend those relations to the complete operation table.

3.3 Model symmetries and the regular representation

Because the three factor families in (1) are untied, the factorization accommodates general binary operations, including arbitrary quasigroups. Crucially, applying independent permutations to the row, column, and output vocabularies corresponds to merely relabeling the parameter slices; i.e., the model is equivariant under isotopy, and the scalar regularizer (3) is inherently isotopy-invariant.

Furthermore, the model possesses a continuous *matrix gauge symmetry*. For any unitary matrices $U, V, W \in \mathrm{U}(n)$, the transformation $(A_a, B_b, C_c) \mapsto (UA_aV^\dagger, VB_bW^\dagger, WC_cU^\dagger)$ leaves both the labeled prediction tensor and the regularizer unchanged.

Finally, for any group target, its left regular representation $\rho_{\mathrm{reg}}(g)$ is unitary with trace $n\mathbf{1}_{\{g=e\}}$. Consequently, the parameter state $\Theta_{\mathrm{reg}} = (\rho_{\mathrm{reg}}(\cdot), \rho_{\mathrm{reg}}(\cdot), \rho_{\mathrm{reg}}(\cdot)^\dagger)$ satisfies the exact reconstruction constraint via:

$$T_{abc}(\Theta_{\mathrm{reg}}) = \frac{1}{n}\operatorname{Tr}\big(\rho_{\mathrm{reg}}(a)\rho_{\mathrm{reg}}(b)\rho_{\mathrm{reg}}(c)^\dagger\big) = \frac{1}{n}\operatorname{Tr}\big(\rho_{\mathrm{reg}}(a \circ b \circ c^{-1})\big) = \mathbf{1}_{\{a\circ b=c\}}. \tag{6}$$

4 Identifiability and sample complexity

4.1 Local locking and coverage

Lemma 1 (Matrix AM–GM; Huh et al., 2025)**.** *For all $X, Y, Z \in \mathbb{C}^{n\times n}$, $\|XY\|_F^2 + \|YZ\|_F^2 + \|ZX\|_F^2 \geq 3n\big|\frac{1}{n}\operatorname{Tr}(XYZ)\big|^{4/3}$. If $\frac{1}{n}\operatorname{Tr}(XYZ) = 1$, equality holds if and only if X, Y, Z are unitary and $XYZ = I$.*

Theorem 2 (Observed-edge locking)**.** *If $(a, b) \in S$ and $T_{ab,a\circ b}(\Theta) = 1$, then $r_\Theta(a, b, a \circ b) \geq 3$. Equality holds if and only if $A_a, B_b, C_{a\circ b}$ are unitary and $A_aB_bC_{a\circ b} = I$; in that case $A_aB_b = C_{a\circ b}^\dagger$, $B_bC_{a\circ b} = A_a^\dagger$, and $C_{a\circ b}A_a = B_b^\dagger$ (apply Lemma 1 to $(A_a, B_b, C_{a\circ b})$; Appendix A.1).*

An observation set $S \subseteq Q^2$ has **three-slot coverage** if every $a \in Q$ occurs as a first input, every $b \in Q$ as a second input, and every $c \in Q$ as an output $a \circ b$ of some $(a, b) \in S$; the three conditions ensure that every slice A_a, B_b, and C_c occurs in at least one observed positive triple.

Theorem 3 (Sampled global minimizers are unitary under coverage)**.** *Let $(Q, \circ)$ be a finite group, and suppose that (1) S has three-slot coverage; (2) the constraints include exact fit on the full observed support Ω_S: $T_{abc}(\Theta) = \delta_{abc}$ for all $(a, b, c) \in \Omega_S$; and (3) $\widehat{\Theta}$ is a global minimizer of $\mathscr{H}_S$ over the observed feasible set. Then the minimum value is $3|S|$ and every matrix slice of $\widehat{\Theta}$ is unitary. The same holds for the positive program (P_+) and for every anchored program (Theorem L.1).*

Proof sketch. To establish that the minimum is exactly $3|S|$, we use the regular representation as a witness: its parameter tuple is feasible and evaluates to exactly $\mathscr{H}_S = 3|S|$. Because each observed summand is bounded below by 3 (Theorem 2), *every* constrained global minimizer must saturate this bound on all observed pairs. The Matrix AM–GM equality characterization, combined with three-slot coverage, therefore forces every factor slice of any such minimizer to be unitary (Appendix A.3).

Remark. While the regular representation itself recovers the full table, this theorem guarantees only that an arbitrary global minimizer $\widehat{\Theta}$ has unitary slices. As shown next, coverage alone does not force these slices to synchronize into a single representation, nor does it guarantee the recovery of unobserved entries (Section 4.4).

Theorem 4 (Regularized common-scale locking)**.** *Let $(Q, \circ)$ be a finite group and suppose that $\lambda > 0$, S has three-slot coverage, and $\widehat{\Theta}$ is an exact global minimizer of $\mathcal{J}_{\lambda,\Omega_S}$. Then there is a*

unique $s_\lambda \in (0,1)$ satisfying $s_\lambda^3 + 2\lambda s_\lambda = 1$, and every factor slice has the same Gram scale:

$$\widehat{A}_a \widehat{A}_a^\dagger = \widehat{B}_b \widehat{B}_b^\dagger = \widehat{C}_c \widehat{C}_c^\dagger = s_\lambda^2 I \qquad (a, b, c \in Q).$$

Moreover, every global minimizer satisfies exact local faithfulness on observed pairs: $T_{abc}(\widehat{\Theta}) = 0$ for all $c \neq a \circ b$ with $(a, b) \in S$. The uniformly scaled regular representation $\Theta_{\mathrm{reg}}(s_\lambda) = (s_\lambda \rho_{\mathrm{reg}}(a),\ s_\lambda \rho_{\mathrm{reg}}(b),\ s_\lambda \rho_{\mathrm{reg}}(c)^\dagger)$ attains the global minimum $|S|\phi_\lambda(s_\lambda)$, where $\phi_\lambda(x) = (x^3 - 1)^2 + 3\lambda x^4$, so a global minimizer exists.

The proof bounds each positive term below by the scalar envelope ϕ_λ, uniquely minimized at s_λ, and shows that every global minimizer attains the Matrix AM–GM equality case $\widehat{A}_a \widehat{B}_b \widehat{C}_{a \circ b} = s_\lambda^3 I$ at every observed pair (Appendix A.4). Hence the normalized slices $s_\lambda^{-1}\widehat{A}_a$, $s_\lambda^{-1}\widehat{B}_b$, $s_\lambda^{-1}\widehat{C}_c$ are unitary and satisfy the locking relations of Theorem 2 on S, and every factor slice is invertible; unitary gauge transformations alone cannot change the raw scale s_λ. Extending local faithfulness to unobserved entries requires further relations among observations (Sections 4.2 and 4.5).

Positive-only degeneracy. Without the negative candidate terms, the scaled trivial state $A_a = B_b = C_c = s_\lambda I_n$ minimizes the remaining functional for every support yet never factorizes $(Q, \circ)$ when $n \geq 2$; under $\mathcal{J}_{\lambda,\Omega_S}$ it pays $(n-1)s_\lambda^6$ per observed pair and is not a global minimizer when $(Q, \circ)$ is a group and $S \neq \emptyset$ (Propositions A.4 and A.5). The tuple $A_a = B_b = C_c = I_n$ is optimal for (P_+), which is why the passive theorem of Section 4.5 imposes an anchor.

Coverage rate. Let S_m be the set of distinct pairs observed in m independent uniform draws from Q^2, where $(Q, \circ)$ is a quasigroup. Then $\Pr(S_m \text{ fails three-slot coverage}) \leq 3n(1 - \frac{1}{n})^m \leq 3ne^{-m/n}$, so $m \geq n \log(3n/\delta)$ draws give three-slot coverage with probability at least $1 - \delta$ (Theorem A.7); the proof is a union bound over the $3n$ slot–symbol events, since the row, column, and product of a uniform pair are each uniform in a quasigroup. With Theorems 3 and 4, the same number of draws makes every exact global minimizer unitary, or of common Gram scale $s_\lambda^2 I$ and full rank, with probability at least $1 - \delta$ (Corollaries A.8 and A.9).

Coverage reaches every factor slice, but it does not determine the unobserved products. We next identify observation patterns that turn these local constraints into a global representation.

4.2 Active teaching and presentation scaffolds

In this subsection, *recovery* means that every exact global minimizer $\widehat{\Theta}$ of $\mathcal{J}_{\lambda,\Omega_S}$ satisfies $T_{abc}(\widehat{\Theta}) = s_\lambda^3 \delta_{abc}$ on all n^3 triples, so that its output argmax is correct on every pair. The teacher knows the operation—hence the identity e, a generating set Γ, a presentation $\langle \Gamma \mid R \rangle$, and a spanning tree of the Cayley graph—and selects S; the learner receives only the labeled triples in S and is not told which role each pair plays.

Definition 5 (Generating scaffolds)**.** Let $G = (Q, \circ)$ be a finite group of order n with identity e, generating set $\Gamma \subset Q$, and finite presentation $\langle \Gamma \mid R \rangle$.

(i) *Identity cross:* $S_{\mathrm{cross}}(e) := (\{e\} \times Q) \cup (Q \times \{e\})$, with $|S_{\mathrm{cross}}| = 2n - 1$.

(ii) *Canonical column scaffold:* $S_{\mathrm{gen}}(e, \Gamma) := S_{\mathrm{cross}}(e) \cup \bigcup_{\gamma \in \Gamma} (Q \times \{\gamma\})$, with $|S_{\mathrm{gen}}| \leq (2n - 1) + |\Gamma|(n - 1)$.

(iii) *Minimal presentation scaffold:* $S_{\min}(e, \Gamma, R) := S_{\mathrm{cross}}(e) \cup T_{\mathrm{Cay}} \cup R_{\mathrm{indep}}$, where T_{Cay} is a directed spanning tree of the Cayley graph $\mathrm{Cay}(G, \Gamma)$ rooted at e, with edges (g, s), $s \in \Gamma$, and R_{indep} is the set of pairs outside $S_{\mathrm{cross}} \cup T_{\mathrm{Cay}}$ met when the defining relators R are traced from e as closed paths; a letter s^{-1} is traced as the element s^{-1} and is admissible when $(s^{-1}, s) \in S_{\min}$ (Appendix B.2).

Theorem 6 (Recovery from active scaffolds)**.** *Let $G = (Q, \circ)$ be a finite group of order n with presentation $\langle \Gamma \mid R \rangle$, let $S \supseteq S_{\mathrm{gen}}(e, \Gamma)$ or $S \supseteq S_{\min}(e, \Gamma, R)$, and let $\widehat{\Theta}$ be an exact global minimizer of $\mathcal{J}_{\lambda,\Omega_S}$. Normalize $\widetilde{A} = s_\lambda^{-1}\widehat{A}$, $\widetilde{B} = s_\lambda^{-1}\widehat{B}$, $\widetilde{C} = s_\lambda^{-1}\widehat{C}$, and apply a gauge transformation so that $\widetilde{A}_e = \widetilde{B}_e = I_n$. Then:*

(i) gauge synchronization: $\widetilde{A}_g = \widetilde{B}_g = \widetilde{C}_g^\dagger =: U_g$ *for all $g \in Q$;*

(ii) representation law: $U_a U_b = U_{a \circ b}$ *for all* $a, b \in Q$*;*

(iii) full-table factorization: $T_{abc}(\widehat{\Theta}) = s_\lambda^3 \delta_{abc}$ *on* Q^3.

Query bounds. *Consequently, under an active query regime, the minimum number* m^+ *of observed input pairs required to recover the complete Cayley table satisfies:*

(a) $m^+ \le |S_{\mathrm{gen}}(e, \Gamma)| \le (2n-1) + |\Gamma|(n-1)$*; in particular* $m^+ \le 3n-2$ *for cyclic groups and* $m^+ \le 4n-3$ *for two-generator groups such as* D_{2k} *and* S_k *(Theorem B.4);*

(b) $m^+ \le |S_{\min}| \le (2n-1) + (n-1) + |R_{\mathrm{indep}}| = 3n - 2 + |R_{\mathrm{indep}}|$*, which depends on the chosen presentation through* $|R_{\mathrm{indep}}|$ *(Theorem B.5).*

Proof sketch. By Theorem 4, the normalized slices are unitary and locked on S, and the identity cross synchronizes the three families after gauge normalization. Generator columns extend $U_g U_s = U_{g \circ s}$ to all pairs by word induction; in $S_{\min}$, the tree names every element as a word in the generators and the traced relator cycles impose the defining relations, so von Dyck's theorem (Lyndon & Schupp, 1977) gives the representation law. Zero negative-target error at an observed pair supplies an anchor, and Theorem 7 yields (iii). Full statements and proofs are Theorems B.4 and B.5; Lean declarations are in Appendix M. □

The presentation scaffold replaces full generator columns by a spanning tree that names every element and by relator traces, so its size grows with $|\Gamma|$ only through $|R_{\mathrm{indep}}|$; constructions for the benchmark families are in Appendix G. Appendix E compares $|S_{\min}|$ with a Fano lower bound for target-independent observations, which does not constrain a target-aware teacher. These results concern structural identifiability of exact global minimizers; they do not assert that gradient-based optimization from random initialization reaches such a minimizer.

4.3 Character anchor

Theorem 6 obtains the homomorphism law on all elements from positive relations; the remaining question is how a learner can enforce character trace orthogonality without observing negative candidate entries across the entire table.

Theorem 7 (Character anchor: one complete output fiber forces the regular representation; Huh et al. (2025))**.** *Let* $R : G \to \mathrm{U}(n)$ *be a unitary representation of a finite group* G*, so that* $R_a R_b = R_{ab}$ *for all* $a, b \in G$*. Suppose that for* ***one observed input pair*** $(a_0, b_0) \in G^2$ *with product* $g_0 = a_0 b_0$*, its complete one-hot output fiber is enforced:* $\frac{1}{n}\mathrm{Tr}(R_{a_0 b_0} R_c^\dagger) = \mathbf{1}_{\{g_0 = c\}}$ *for all* $c \in G$*. Then (i)* $\chi_R(h) := \mathrm{Tr}(R_h) = n \mathbf{1}_{\{h=e\}}$ *for all* $h \in G$*; (ii) there exists* $U \in \mathrm{U}(n)$ *with* $R_g = U \rho_{\mathrm{reg}}(g) U^\dagger$ *for all* g*; and (iii)* $T_R(a, b, c) := \frac{1}{n}\mathrm{Tr}(R_a R_b R_c^\dagger) = \delta_{abc}$ *on* ***all*** n^3 *triples, observed and unobserved alike.*

Setting $c = h^{-1} g_0$ gives $R_{g_0} R_c^\dagger = R_h$, and $c = g_0$ exactly when $h = e$, so the single fiber reads off the regular character (i). Part (iii) follows from (i), since $\frac{1}{n}\mathrm{Tr}(R_a R_b R_c^\dagger) = \frac{1}{n}\mathrm{Tr}(R_{abc^{-1}})$, and (ii) follows because a representation is determined up to (unitary) equivalence by its character (Serre, 1977) (Appendix B.3). A learner therefore does not require negative samples across all n^3 triples or across all $|S|n$ candidates: one complete one-hot output fiber, which costs $n-1$ scalar zero constraints $\frac{1}{n}\mathrm{Tr}(R_{g_0} R_c^\dagger) = 0$ for $c \neq g_0$, is a sufficient negative-entry budget to eliminate the trivial representation and all non-regular representations.

An active teacher can select relations that synchronize the factors and determine the representation. Passive sampling must produce sufficient relations without a prescribed scaffold; connectivity of the observed input-pair graph alone does not provide them.

4.4 Why coverage and connectivity do not suffice

Proposition 8 (Ambiguity of the identity cross)**.** *An arbitrary spanning tree of the input-pair graph need not even synchronize the factor families, despite three-slot coverage. Appendix C.1 gives a counterexample on* C_3*: an exact global minimizer whose normalized tuple satisfies all observed full-output constraints, yet retains a sign ambiguity that prevents synchronization and causes an incorrect prediction at an unobserved pair. The comparison graphs introduced in Section 4.5 encode*

equalities obtained from pairs of observed products; their connectivity allows these equalities to propagate and establish a common representation.

4.5 PASSIVE BERNOULLI RECOVERY

We outline the comparison-graph argument and state the resulting recovery guarantee here; detailed statements and proofs are given in Appendix L. Let $\Omega \subseteq G^2$ be a Bernoulli-p mask, which includes each input pair independently with probability p; Ω plays the role of S, and $\bar{m} = \mathbb{E}|\Omega| = pn^2$. By Theorem L.1, every global minimizer of (P_+) on Ω has unitary slices on observed triples and satisfies $A_a B_b = D_{ab}$ with $D_c := C_c^\dagger$ at every observed pair.

Two observations that share a column give one **relative comparison** (Lemma L.3). If $A, B, D : G \to \mathrm{U}(d)$ satisfy the observed laws and $(a, b), (ga, b) \in \Omega$, associativity and cancellation of the common right factor give

$$D_{g(ab)} D_{ab}^\dagger = (A_{ga} B_b)(A_a B_b)^\dagger = A_{ga} A_a^\dagger.$$

Thus the relative matrices $P_g(a) := A_{ga} A_a^\dagger$ and $Q_g(x) := D_{gx} D_x^\dagger$ satisfy $P_g(a) = Q_g(ab)$ (Figure 1a,b). This compares the two left translations without assuming either unobserved multiplication law or identifying the factor families.

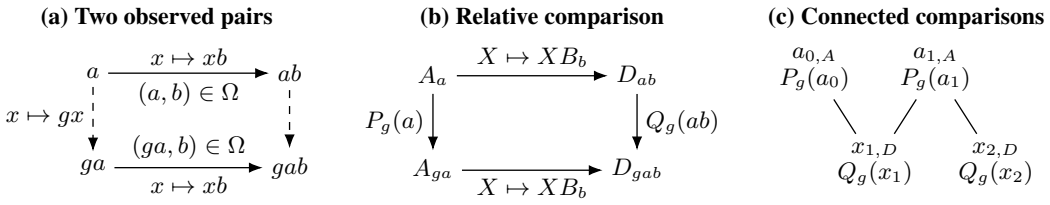


Figure 1: From observed products to common left multipliers. (a) The label square commutes by associativity; dashed sides assume no observed multiplication law. (b) Vertical arrows apply the indicated left multipliers, which agree after canceling B_b. (c) Undirected edges propagate equality: $P_g(a_0) = Q_g(x_1) = P_g(a_1) = Q_g(x_2)$.

For $g \neq e$, the **relative-comparison graph** $\mathcal{K}_g(\Omega)$ has vertex classes G_A, G_D. Vertex $a \in G_A$, carrying $P_g(a)$, is adjacent to $x \in G_D$, carrying $Q_g(x)$, exactly when $(a, a^{-1}x), (ga, a^{-1}x) \in \Omega$. Each edge records equality, so connected paths give one common left multiplier U_g (Figure 1c). The mask is **good**, written $\mathcal{E}_G(\Omega)$, if it has three-slot coverage and all these graphs are connected; this property persists when observations are added (Definitions L.4 and L.5).

These common left multipliers satisfy $U_g U_h = U_{gh}$. Column coverage then determines the remaining B-family, synchronizing the three families (Theorem L.8); one complete output fiber selects the regular character. The resulting *regular-gauge tuple* (L.2) has the form $A_a = X \rho_{\mathrm{reg}}(a) Y^\dagger$, $B_b = Y \rho_{\mathrm{reg}}(b) Z^\dagger$, $C_c = Z \rho_{\mathrm{reg}}(c)^\dagger X^\dagger$ for fixed $X, Y, Z \in \mathrm{U}(n)$, and recovers the labeled tensor.

Each comparison edge requires two distinct observed cells, so its probability is p^2 and each vertex has expected degree np^2. Edges can share cells; bounded overlap and a union bound over cuts and translations give the sufficient condition below (Lemmas L.12 and L.13 and Proposition L.15). With $\bar{m} = pn^2$, this yields the order $\boxed{\mathcal{O}(n^{3/2}\sqrt{\log n})}$. Appendix L.5 motivates the relative-comparison argument through associativity squares, explaining how two-observation comparisons replace three-observation certificates for individual missing laws.

$$\boxed{p^2 n \geq 6 \log \frac{8n^2}{\delta}} \tag{7}$$

Theorem 9 (Exact partial recovery without shared embedding)**.** *Let G be a fixed finite group of order $n \geq 2$, let $0 < \delta < 1$, and let $\Omega \subseteq G^2$ be a Bernoulli-p mask. Use the three-family model* (1)*, the masked regularizer* (3)*, and $n \times n$ complex slices. Then, for every mask, the positive constrained program (P_+) has minimum $3|\Omega|$. If* (7) *holds, then there is an event of probability at least $1 - \delta$ on which the following statements hold simultaneously:*

1. *Every positive-only global minimizer has unitary slices and is of the form $A_a = U_a P,\ B_b = P^\dagger U_b Q,\ C_c = Q^\dagger U_c^\dagger$, for a representation $U : G \to \mathrm{U}(n)$ and fixed unitaries P, Q. In particular, $A_a B_b = C_{ab}^\dagger$ at every input pair.*
2. *The mask is nonempty. At any observed pair selected after seeing the mask, adding its full-output constraints leaves the minimum at $3|\Omega|$ and forces every augmented-program global minimizer to be a regular-gauge tuple.*
3. *The anchored minimizers recover the entire labeled tensor: $T_{a,b,c}(\Theta) = \mathbf{1}_{\{ab=c\}}$ for all triples. The same conclusion holds when full-output constraints are imposed at any nonempty set of observed pairs, in particular at all of them.*

Corollary 10 (Expected pair count)**.** *Condition* (7) *is equivalent to $\bar{m} = pn^2 \geq \sqrt{6}\, n^{3/2}(\log(8n^2/\delta))^{1/2}$, a sufficient order* $\boxed{\mathcal{O}(n^{3/2}\sqrt{\log n})}$ *in expected observed pairs for fixed confidence (Appendix L.4).*

Corollary 11 (Finite-regularization transfer)**.** *Let $\lambda > 0$. On the event $\mathcal{E}_G(\Omega)$, every exact global minimizer $\widehat{\Theta}$ of $\mathcal{J}_{\lambda,\Omega}$ satisfies $T_{\widehat{\Theta}}(a,b,c) = s_\lambda^3 \mathbf{1}_{\{ab=c\}}$ for all $(a,b,c) \in G^3$; in particular its output argmax is correct on every pair. Under* (7) *this holds with probability at least $1-\delta$, simultaneously for all such minimizers (proof in Appendix L.4). The same holds for the fixed-size masks of Section 5, with probability at least $1-2\delta$ (Corollary L.16).*

Thus, for group and group-isotope targets, the HyperCube model with the full-output objective (5) achieves complete-table recovery at every exact global optimum, with high probability, from subquadratically many randomly observed input pairs. This provides a sample-complexity guarantee for generalization from partial observations.

This sufficient bound applies simultaneously to all global minimizers on the stated event. We next compare its dependence on group order with recovery thresholds measured under finite training budgets.

5 Empirical results

Protocol. We train the three-family model with real-valued $n \times n$ slices by minimizing $\mathcal{J}_{\lambda,\Omega_S}$ of (5) with sum reduction and $\lambda = 0.5$, using Adam (learning rate 0.05) for at most 2000 epochs, with early stopping rules that read the unobserved accuracy (initialization, optimizer settings, and stopping rules in Appendix K). The benchmark contains **32 groups** of orders $n \in [4, 128]$ from four families—cyclic, dihedral, permutation (symmetric and alternating), and elementary Abelian $(C_2)^k$—including S_5 and $(C_2)^7$ (Appendix I). For each group, S is a uniformly random set of m distinct input pairs, with m the nearest integer to $c\, n^{1.15} \ln n$ for four to six per-group values of c; each condition uses 5 seeds, and a seed fixes both the mask and the initialization. Accuracy is argmax accuracy on unobserved pairs. We report $m^*(95\%)$, the smallest tested m with mean unobserved accuracy $\geq 95\%$, and $m^*(100\%)$, the smallest tested m at which all five runs reach $\geq 99.9\%$ unobserved accuracy (`results/passive_sampling_results.json`).

Scaling across 32 groups. The measured thresholds $m^*(95\%)$ are well described by the empirical fit

$$m_{\mathrm{hybrid}}(n) = \max\Big\{3n,\ n\ln(3n),\ 0.476\, n^{3/2}\sqrt{\ln n}\Big\},$$

with only the coefficient 0.476 fitted ($R^2 = 0.998$). The log–log slope of $m^*(95\%)$ is 1.401 ± 0.044 over all 32 groups and 1.633 ± 0.031 over the 18 groups with $n \geq 24$, and the observed fraction at $m^*(95\%)$ falls from 68.8% for C_4 to 9.4% for $(C_2)^7$ (Table 3).

Normalizing by $m_{\mathrm{hybrid}}(n)$ reveals similar accuracy transitions across the tested group families: over the 168 tested conditions, mean unobserved accuracy was below 95% whenever $m/m_{\mathrm{hybrid}} < 0.90$ and at least 95% whenever $m/m_{\mathrm{hybrid}} > 1.10$.

Interpretation. The theory explains why observing only part of a group table can be enough: relations between matrix slices allow the observed entries to determine the unobserved ones, making recovery possible with far fewer than n^2 samples. The experiments show a corresponding practical trend: over the tested range, larger groups reach high unobserved accuracy from a smaller fraction

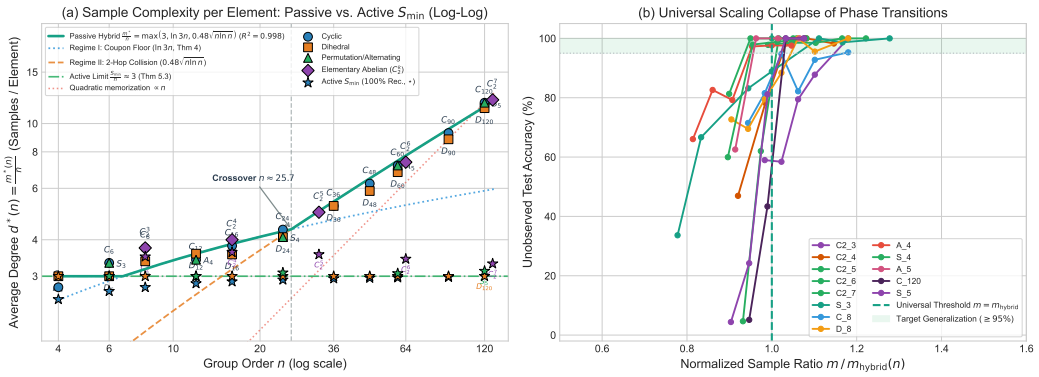


Figure 2: Passive thresholds and active scaffolds across the benchmark. (a) Measured passive threshold per element, $m^*(95\%)/n$, by group family (log–log), with the hybrid fit $m_{\text{hybrid}}(n)/n = \max\{3, \ln(3n), 0.476\sqrt{n \ln n}\}$; stars mark the presentation-scaffold size $|S_{\min}|/n$. (b) Unobserved accuracy versus the normalized ratio $m/m_{\text{hybrid}}(n)$ for a subset of the groups; the dashed line marks $m = m_{\text{hybrid}}$, and the band marks accuracy $\geq 95\%$.

of the table. This suggests that the sample efficiency established theoretically for global minimizers can also be realized in practice through training. These results support the theory's broad message of sample-efficient recovery. The experiments demonstrate high average accuracy after training; the theorem guarantees complete recovery at every global optimum under its stated conditions (Theorem 9 and Appendix I).

Active teaching. On S_5 ($n = 120$), the identity cross (239 pairs) stays at chance on unobserved pairs and adding one generator column does not recover the table, whereas the presentation scaffold (371 pairs; 375 with the current constructor) recovers it in every run (Table 5). The 371-pair scaffold uses 2.58% of the table, against 9.8% for passive recovery of S_5 (Table 3). Appendix J reports the full S_5 results, the same comparison on S_4 (Figure 4), and scaffold ablations for $n = 8$ (Figure 3).

6 Conclusion

We studied the recovery of finite operation tables from partial observations as a canonical setting for algebraic generalization. Gradient sensitivity regularization locks every covered slice to a common scaled-unitary geometry through the equality case of the Matrix AM–GM inequality, which requires only three-slot coverage. Recovering unobserved entries then requires sufficient algebraic relations among the observed products. Under active query selection, a teacher can deterministically scaffold the entire table using only $3n - 2 + |R_{\text{indep}}|$ queries derived from the group's presentation relators. Under passive uncoordinated sampling, connected comparison graphs propagate the associativity prior globally; crucially, this passive bound establishes that exact recovery is guaranteed from a vanishing fraction of the table asymptotically. Measured thresholds across 32 groups are consistent with growth of order $n^{3/2}$ up to logarithmic factors, the order of this passive bound.

Open questions. Three questions remain. *Finite-size gap:* closing the gap between the subquadratic sufficient condition of Theorem 9 and the measured finite-size recovery thresholds remains an open problem. *Necessity:* a label that never occurs as a first (or second) input leaves its A- (or B-) slice absent from the objective and from every constraint, and the n rows of a Bernoulli-p mask are independent, so recovery with probability at least $1 - \delta$ at fixed $\delta < 1$ requires $\bar{m} \geq (1 - o(1))\, n \log n$; whether order $n^{3/2}(\log n)^{1/2}$ is asymptotically necessary is open. *Optimization:* whether gradient-based training reaches the characterized minimizers at these sample sizes.

AI use statement

We used generative AI tools for three purposes. First, to help translate our proofs into the Lean 4 formalization described in Appendix M. Second, to polish the language, editing, and formatting of parts of the manuscript; these edits were reviewed by the authors and introduced no technical

content. Third, to generate the diagrams in Figure 1. All research ideas, mathematical claims and proofs, and the numerical experiment code and plots in this paper were developed by the authors without generative AI. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is theoretical. Its experiments use synthetic operation tables of finite groups generated from standard group presentations; they involve no human subjects, personal data, or deployed systems. We are not aware of ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

Complete statements and proofs of all results are given in Appendices A to E and L, and the Lean 4 sources, the correspondence between manuscript statements and formal declarations, and the axiom audit are available in the anonymous repository described in Appendix M. The benchmark groups, presentations, and scaffold constructions are specified in Appendices G and I; the objective, initialization, optimizer settings, and stopping rules are given in Appendix K; and each reported value is traced to a named result file in Section 5 and Appendices I and J.

REFERENCES

László Babai, Albert J. Goodman, William M. Kantor, Eugene M. Luks, and Péter P. Pálfy. Short presentations for finite groups. *Journal of Algebra*, 194(1):79–112, 1997.

Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-testing/correcting with applications to numerical problems. *Journal of Computer and System Sciences*, 47(3):549–595, 1993.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006.

Funda Ergün, Sampath Kannan, S. Ravi Kumar, Ronitt Rubinfeld, and Mahesh Viswanathan. Spot-checkers. *Journal of Computer and System Sciences*, 60(3):717–751, 2000.

Michael P. Friedlander, Halyun Jeong, Yaniv Plan, and Özgür Yılmaz. NBIHT: An efficient algorithm for 1-bit compressed sensing with optimal error decay rate. *IEEE Transactions on Information Theory*, 68(2):1157–1177, 2022. doi: 10.1109/TIT.2021.3124598. URL https://doi.org/10.1109/TIT.2021.3124598.

Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Sally A. Goldman and Michael J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995.

Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023.

Dongsung Huh. Discovering group structures via unitary representation learning. In *International Conference on Learning Representations*, 2025.

Dongsung Huh, Lior Horesh, and Halyun Jeong. A differentiable measure of algebraic complexity: Provably exact discovery of group structures. *arXiv preprint arXiv:2511.23152*, 2025. URL https://arxiv.org/abs/2511.23152v4. Version 4, revised 19 May 2026.

R. Kaas and J. M. Buhrman. Mean, median and mode in binomial distributions. *Statistica Neerlandica*, 34(1):13–18, 1980.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Roger C. Lyndon and Paul E. Schupp. *Combinatorial Group Theory*. Springer, 1977.

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J. Sutherland. Why do you grok? a theoretical analysis on grokking modular addition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35934–35967, 2024. URL https://proceedings.mlr.press/v235/mohamadi24a.html.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

László Pyber. Enumerating finite groups of given order. *Annals of Mathematics*, 137(1):203–220, 1993.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Jean-Pierre Serre. *Linear Representations of Finite Groups*, volume 42 of *Graduate Texts in Mathematics*. Springer, 1977.

J. H. van Lint and R. M. Wilson. *A Course in Combinatorics*. Cambridge University Press, second edition, 2001.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A PROOFS OF METRIC LOCKING, REGULARIZED SCALE, AND COVERAGE

Definition A.1 (Three-slot coverage)**.** An observation set $S \subseteq Q^2$ has three-slot coverage if every $a \in Q$ occurs as a first input, every $b \in Q$ as a second input, and every $c \in Q$ as an output $a \circ b$ of some $(a, b) \in S$. Equivalently, every slice A_a, B_b, and C_c occurs in at least one observed positive triple.

A.1 MATRIX AM–GM

The scalar envelope and common-scale geometry rely on the Matrix AM–GM inequality evaluated at an arbitrary positive prediction scale.

Lemma A.2 (Matrix AM–GM)**.** *For any matrices $X, Y, Z \in \mathbb{C}^{n\times n}$,*

$$\frac{1}{n}\left(\|XY\|_F^2 + \|YZ\|_F^2 + \|ZX\|_F^2\right) \geq 3\left|\frac{1}{n}\operatorname{Tr}(XYZ)\right|^{4/3}.$$

In particular, if $\frac{1}{n}\operatorname{Tr}(XYZ) = s^3$ with $s > 0$, then $\frac{1}{n}\left(\|XY\|_F^2 + \|YZ\|_F^2 + \|ZX\|_F^2\right) \geq 3s^4$, with equality if and only if $XX^\dagger = YY^\dagger = ZZ^\dagger = s^2 I$ and $XYZ = s^3 I$.

Proof. The inequality is Lemma 1 divided by n. For the scaled statement, define $\widetilde{X} = X/s$, $\widetilde{Y} = Y/s$, $\widetilde{Z} = Z/s$. Then $\frac{1}{n}\operatorname{Tr}(\widetilde{X}\widetilde{Y}\widetilde{Z}) = 1$, and Lemma 1 (`lemma16_matrix_amgm`) gives $\frac{1}{n}\left(\|\widetilde{X}\widetilde{Y}\|_F^2 + \|\widetilde{Y}\widetilde{Z}\|_F^2 + \|\widetilde{Z}\widetilde{X}\|_F^2\right) \geq 3$. Since $\|\widetilde{X}\widetilde{Y}\|_F^2 = s^{-4}\|XY\|_F^2$, multiplying by s^4 recovers the scaled inequality. Equality in Lemma 1 forces $\widetilde{X}, \widetilde{Y}, \widetilde{Z}$ to be unitary with product I, which yields $XX^\dagger = YY^\dagger = ZZ^\dagger = s^2 I$ and $XYZ = s^3 I$ (`matrix_amgm_at_positive_scale_equality` in `HyperCubeGroup/SampleComplexity/RegularizedObservedSupport.lean`). □

Proof of Theorem 2. Apply Lemma 1 with $(X, Y, Z) = (A_a, B_b, C_{a\circ b})$ and $\frac{1}{n}\operatorname{Tr}(A_a B_b C_{a\circ b}) = 1$, and divide by n: $h_\Theta(a, b, a \circ b) \geq 3$, i.e. $r_\Theta(a, b, a \circ b) \geq 3/n$. Equality holds if and only if the three slices are unitary and $A_a B_b C_{a\circ b} = I$. Multiplying this identity on the right by $C_{a\circ b}^\dagger$, on the left by $A_a^\dagger$, and cyclically rearranging $C_{a\circ b} A_a B_b = I$ give the three locked relations (`local_locking_of_tripleSensitivity_eq_three`). □

A.2 SCALE ENVELOPE UNIQUENESS AND EXACT POLYNOMIAL FACTORIZATION

The one-dimensional scalar objective along a common-scale unitary triple is $\phi_\lambda(x) = (x^3 - 1)^2 + 3\lambda x^4$.

Lemma A.3 (Scale envelope unique minimizer)**.** *For any $\lambda > 0$, the equation $s^3 + 2\lambda s = 1$ has a unique positive root $s_\lambda \in (0, 1)$, and s_λ is the unique global minimizer of $\phi_\lambda(x)$ over $x \geq 0$.*

Proof. The first derivative is $\phi'_\lambda(x) = 6x^2(x^3 - 1) + 12\lambda x^3 = 6x^2(x^3 + 2\lambda x - 1)$. The function $g(x) = x^3 + 2\lambda x$ is strictly increasing on $[0, \infty)$ with $g(0) = 0$ and $g(1) = 1 + 2\lambda > 1$. By the intermediate value theorem, there is a unique $s_\lambda \in (0, 1)$ satisfying $g(s_\lambda) = 1$. To establish global minimality, we factor the exact difference using the stationarity identity $s_\lambda^3 + 2\lambda s_\lambda = 1$:

$$2s_\lambda(\phi_\lambda(x) - \phi_\lambda(s_\lambda)) = (x - s_\lambda)^2 Q(s_\lambda, x),$$

$$Q(s_\lambda, x) = 2s_\lambda x^4 + 4s_\lambda^2 x^3 + (3 + 3s_\lambda^3)x^2 + (2s_\lambda + 2s_\lambda^4)x + (s_\lambda^2 + s_\lambda^5).$$

Since $s_\lambda > 0$ and $x \geq 0$, every coefficient of $Q(s_\lambda, \cdot)$ is positive, so $Q(s_\lambda, x) > 0$ for all $x \geq 0$. Hence $\phi_\lambda(x) > \phi_\lambda(s_\lambda)$ for all $x \neq s_\lambda$ (`scaleEnvelope_gap_factorization` and `scaleEnvelope_unique_minimizer` in `HyperCubeGroup/SampleComplexity/RegularizedObservedSupport.lean`). □

A.3 Proof of global slice unitarity under coverage

Proof of Theorem 3. The regular representation $\Theta_{\mathrm{reg}} = (\rho_{\mathrm{reg}}(a), \rho_{\mathrm{reg}}(b), \rho_{\mathrm{reg}}(c)^\dagger)$ of the group is feasible for the complete operation tensor by (6), hence for every observed subset Ω_S ($T_{ab,a\circ b}(\Theta_{\mathrm{reg}}) = 1$ and $T_{ab,c}(\Theta_{\mathrm{reg}}) = 0$ for $c \neq a \circ b$). Every active slice in this certificate is unitary, so each observed term equals 3 and $\inf_{\Theta \text{ observed-feasible}} \mathscr{H}_S(\Theta) \leq 3|S|$. By Theorem 2, for every observed-feasible parameter state Θ and each $(a, b) \in S$, $h_\Theta(a, b, a \circ b) \geq 3$. Summing over S gives $\mathscr{H}_S(\Theta) \geq 3|S|$. Therefore $\mathscr{H}_S(\widehat{\Theta}) = 3|S|$. Since each summand is at least 3 and the sum attains $3|S|$, every observed transition saturates its local lower bound. By the equality condition of Theorem 2, the three active slices $A_a, B_b, C_{a\circ b}$ on each observed edge are unitary. Three-slot coverage ensures that each slice A_a, B_b, C_c appears in at least one observed transition, so every factor slice of $\widehat{\Theta}$ is unitary (`constrained_support_global_min_unitary_of_threeSlotCoverage`). The same argument applies verbatim to (P_+) and to every anchored program, because the certificate satisfies all of their constraints (Theorem L.1 and Corollary L.2). □

A.4 Proof of regularized common-scale locking

Proof of Theorem 4. Decompose $\mathcal{J}_{\lambda,\Omega_S}$ into positive transitions and negative candidate outputs:

$$\mathcal{J}_{\lambda,\Omega_S}(\Theta) = \sum_{(a,b)\in S} \left[|1 - T_{ab,a\circ b}|^2 + \lambda h_\Theta(a, b, a \circ b)\right] + \sum_{(a,b)\in S} \sum_{c\neq a\circ b} |T_{abc}|^2.$$

Fix $(a, b) \in S$, let $t = T_{ab,a\circ b}(\Theta)$, and put $x = |t|^{1/3} \geq 0$. The reverse triangle inequality gives $|1-t|^2 \geq (1-|t|)^2 = (x^3-1)^2$, and Lemma A.2 gives $h_\Theta(a, b, a\circ b) \geq 3|t|^{4/3} = 3x^4$. Hence each positive term is at least $\phi_\lambda(x) \geq \phi_\lambda(s_\lambda)$ by Lemma A.3. Since the negative sum is nonnegative, $\mathcal{J}_{\lambda,\Omega_S}(\Theta) \geq |S|\phi_\lambda(s_\lambda)$ for all Θ.

The uniformly scaled regular representation $\Theta_{\mathrm{reg}}(s_\lambda) = (s_\lambda\rho_{\mathrm{reg}}(a), s_\lambda\rho_{\mathrm{reg}}(b), s_\lambda\rho_{\mathrm{reg}}(c)^\dagger)$ attains this lower bound. By (6), $T_{ab,a\circ b} = s_\lambda^3$ and $T_{abc} = 0$ for all $c \neq a \circ b$, and $h = 3s_\lambda^4$, so every observed pair contributes $\phi_\lambda(s_\lambda)$ and the negative sum vanishes. Hence $\inf_\Theta \mathcal{J}_{\lambda,\Omega_S}(\Theta) = |S|\phi_\lambda(s_\lambda)$, attained (`scaledLeftRegularParams_is_global_min`).

Every global minimizer therefore makes each inequality above an equality at every observed pair: (i) $x = s_\lambda$, by uniqueness in Lemma A.3; (ii) $|1 - t| = 1 - |t|$ with $|t| = s_\lambda^3 < 1$, so $t = s_\lambda^3$ is real and positive; (iii) equality in Lemma A.2 at scale s_λ, so $A_aA_a^\dagger = B_bB_b^\dagger = C_{a\circ b}C_{a\circ b}^\dagger = s_\lambda^2 I$ and $A_aB_bC_{a\circ b} = s_\lambda^3 I$; and (iv) zero negative error, $T_{abc}(\widehat{\Theta}) = 0$ for all $c \neq a \circ b$ on observed pairs. Three-slot coverage then reaches all slices (`global_min_exists_common_gram_scale`, `categorical_global_min_common_gram_scale`). □

A.5 Positive-only degeneracy and the role of negative targets

This subsection collects the complete statements behind Proposition A.4 and the characterization of the negative-target error; it merges the nonduplicated content of the draft's two propositions on contrastive supervision.

Proposition A.4 (Positive-only degeneracy: necessity of negative targets)**.** *Let $(Q, \circ)$ be any binary operation of order $n \geq 2$, and let $S \subseteq Q \times Q$ be any observation support. Omitting the negative candidate terms from $\mathcal{J}_{\lambda,\Omega_S}$ yields the positive-only functional $\mathcal{J}^+_{\lambda,S}(\Theta) := \sum_{(a,b)\in S}[|1 - T_{ab,a\circ b}|^2 + \lambda h_\Theta(a, b, a \circ b)]$. Then:*

1. ***Trivial minimizer on positive-only loss.*** *The scaled trivial state $\Theta_{triv}(s_\lambda)$, $A_a = B_b = C_c = s_\lambda I_n$, satisfies the common Gram condition $MM^\dagger = s_\lambda^2 I$ and attains the scalar envelope value $\phi_\lambda(s_\lambda)$ on every observed edge. Consequently it is an exact global minimizer: $\mathcal{J}^+_{\lambda,S}(\Theta_{triv}(s_\lambda)) = |S|\phi_\lambda(s_\lambda) = \inf_\Theta \mathcal{J}^+_{\lambda,S}(\Theta)$.*
2. ***Failure of factorization.*** *For any $s > 0$, $\Theta_{triv}(s)$ never factorizes $(Q, \circ)$: for every pair (a, b) and every incorrect candidate $c \neq a \circ b$, $T_{abc}(\Theta_{triv}) = s^3 > 0 = \delta_{abc}$.*
3. ***Sub-optimality under the full objective.*** *If in addition $S \neq \emptyset$ and $(Q, \circ)$ is a finite group, so that $\inf_\Theta \mathcal{J}_{\lambda,\Omega_S} = |S|\phi_\lambda(s_\lambda)$ (the first two paragraphs of the proof of Theorem 4 do not use*

coverage), the trivial state incurs the contrastive penalty $(n-1)s_\lambda^6$ per observed pair and is not a global minimizer: $\mathcal{J}_{\lambda,\Omega_S}(\Theta_{triv}(s_\lambda)) = |S|\phi_\lambda(s_\lambda) + |S|(n-1)s_\lambda^6 > \inf_\Theta \mathcal{J}_{\lambda,\Omega_S}(\Theta)$.

Proof. (1) Direct evaluation gives $T_{abc}(\Theta_{\text{triv}}(s)) = \frac{1}{n}\operatorname{Tr}(s^3 I_n) = s^3$ for every triple and $h = \frac{1}{n}\cdot 3\left\|s^2 I_n\right\|_F^2 = 3s^4$, so each observed term equals $\phi_\lambda(s)$, which is $\phi_\lambda(s_\lambda)$ at $s = s_\lambda$. The lower bound $\mathcal{J}^+_{\lambda,S} \geq |S|\phi_\lambda(s_\lambda)$ is the positive-term part of the proof of Theorem 4, which uses no property of $(Q,\circ)$. (2) Immediate from $T_{abc} = s^3$. (3) Each observed pair has exactly $n-1$ incorrect candidates, each contributing $|0 - s_\lambda^3|^2 = s_\lambda^6$ (`scaledTrivialParams_is_global_min`, `not_factorizes_scaledTrivialParams`, `scaledTrivialParams_not_categorical_global_min`, `scaledTrivialParams_negativeMse_pos`). □

Proposition A.5 (Zero set of the negative-target error on a representation)**.** *Let S have three-slot coverage, and let Θ have slices $A_g = \rho(g)$, $B_g = \rho(g)$, $C_g = \rho(g)^\dagger$ for a unitary representation $\rho : G \to \mathrm{U}(n)$, so that $T_{abc}(\Theta) = \frac{1}{n}\operatorname{Tr}(\rho(a\circ b)\rho(c)^\dagger)$. Define $\mathcal{L}_{neg}(\Theta) := \frac{1}{|S|}\sum_{(a,b)\in S}\sum_{c\neq a\circ b}|T_{abc}(\Theta)|^2$. Then $\mathcal{L}_{neg}(\Theta) = 0$ if and only if $\frac{1}{n}\operatorname{Tr}(\rho(g)\rho(h)^\dagger) = 0$ for all distinct $g \neq h \in Q$. In that case ρ is faithful, and its character is the regular character $n\mathbf{1}_{\{g=e\}}$.*

Proof. For $(a,b) \in S$ with product $g = a\circ b$, the negative error at (a,b) is $\sum_{c\neq g}|\frac{1}{n}\operatorname{Tr}(\rho(g)\rho(c)^\dagger)|^2$, a sum of nonnegative terms. By output coverage every $g \in Q$ is the product of some observed pair, so $\mathcal{L}_{\text{neg}} = 0$ exactly when all the stated traces vanish (`negativeMse_eq_zero_of_traceOrthogonal`). If $\rho(g) = \rho(h)$ for some $g \neq h$, then $\frac{1}{n}\operatorname{Tr}(\rho(g)\rho(h)^\dagger) = 1$, so ρ is faithful. Finally $\operatorname{Tr}\rho(gh^{-1}) = \operatorname{Tr}(\rho(g)\rho(h)^\dagger) = 0$ for $g \neq h$ shows $\chi_\rho(k) = 0$ for $k \neq e$, while $\chi_\rho(e) = n$. Theorem 7 shows that the same conclusion already follows from zero negative error at a single observed pair. □

A.6 Full rank and the sensitivity identity

Corollary A.6 (Full rank and exact sensitivity identity)**.** *Under the assumptions of Theorem 4, every factor slice of $\widehat{\Theta}$ is invertible. Moreover, the common-scale condition gives the exact identity*

$$|S|\sum_{a,b,c\in Q} h_{\widehat{\Theta}}(a,b,c) = n^3 \sum_{(a,b)\in S} h_{\widehat{\Theta}}(a,b,a\circ b).$$

Proof. Invertibility follows from $MM^\dagger = s_\lambda^2 I$ with $s_\lambda > 0$ (`globalMin_fullRankSlices_of_threeSlotCoverage`). For any two slices X, Y with Gram matrices $s_\lambda^2 I$, $\|XY\|_F^2 = \operatorname{Tr}(XYY^\dagger X^\dagger) = s_\lambda^4 n$, so $h_{\widehat{\Theta}}(a,b,c) = 3s_\lambda^4$ for every triple. Both sides equal $3s_\lambda^4 n^3|S|$. □

This equality is specific to exact global minimizers with the common scaled-unitary Gram structure. It is not a domination bound for arbitrary parameter states on a partial support.

A.7 Proof of the three-slot coverage bound

Theorem A.7 (Three-slot coverage with high probability)**.** *Let $(Q,\circ)$ be a quasigroup of order n, and let S_m be the set of distinct pairs observed in m independent uniform draws from Q^2. Then $\Pr(S_m$ fails three-slot coverage$) \leq 3n\left(1-\frac{1}{n}\right)^m \leq 3ne^{-m/n}$. Consequently, for any $\delta \in (0,1)$, $m \geq n\log(3n/\delta)$ implies three-slot coverage with probability at least $1-\delta$.*

Proof. For a fixed element $q \in Q$ and slot $k \in \{1,2,3\}$, let $E_{q,k}$ be the event that q does not appear in slot k across all m draws. In a quasigroup, the row A_t, the column B_t, and the product $C_t = A_t \circ B_t$ of a uniform pair are each uniformly distributed over Q. Hence each draw places q in slot k with probability $1/n$, and independence of the draws gives $\Pr(E_{q,k}) = (1-1/n)^m \leq e^{-m/n}$.

A union bound over the $3n$ slot–symbol events yields

$$\Pr(S_m \text{ fails three-slot coverage}) = \Pr\Big(\bigcup_{k=1}^{3}\bigcup_{q\in Q} E_{q,k}\Big) \le \sum_{k=1}^{3}\sum_{q\in Q}\Pr(E_{q,k}) \le 3n\left(1-\frac{1}{n}\right)^m \le 3ne^{-m/n}.$$

Setting $3ne^{-m/n} \le \delta$ and solving for m yields $m \ge n\log(3n/\delta)$. □

Corollary A.8 (Sample complexity for global slice unitarity)**.** *Under the assumptions of Theorem 3 and uniform sampling with replacement, $m \ge n\log(3n/\delta)$ draws ensure that every exact constrained global minimizer has unitary slices with probability at least $1-\delta$. Thus $\mathcal{O}(n\log(n/\delta))$ positive observations suffice for global slice unitarity in the exact-minimization regime.*

Corollary A.9 (Sample complexity for common-scale locking)**.** *Under the assumptions of Theorem 4 and uniform sampling with replacement, $m \ge n\log(3n/\delta)$ draws ensure that every exact global minimizer of $\mathcal{J}_{\lambda,\Omega_{S_m}}$ has common Gram scale $s_\lambda^2 I$ with probability at least $1-\delta$. Thus the same $\mathcal{O}(n\log(n/\delta))$ coverage bound suffices for common-scale locking in the equal-weight, exact-global-minimization regime.*

Both corollaries follow by intersecting the coverage event of Theorem A.7 with the deterministic conclusions of Theorems 3 and 4.

B Proofs of scaffold rigidity and word induction

B.1 Word matrix folding and associativity induction

Let $G = (Q, \circ)$ be a group generated by Γ. We formalize word evaluation and representation induction. Word evaluation is wordEval$(e, w) := w.\text{foldl}(f.\text{op}, e)$ for a list $w \in \Gamma^*$, and the matrix product is wordMatrixProd$(A, w) := w.\text{foldl}(\cdot, I)$, where each step multiplies on the right by A_s.

Lemma B.1 (Left multiplication in foldl)**.** *For any family A and matrix M, $w.\mathit{foldl}((\cdot), M) = M \cdot w.\mathit{foldl}((\cdot), I)$.*

Proof. By induction on the list w. For $w = []$, both sides equal M. For $w = s :: ws$,

$$(s :: ws).\text{foldl}((\cdot), M) = ws.\text{foldl}((\cdot), MA_s) = (MA_s)\cdot ws.\text{foldl}((\cdot), I) = M\cdot\big(A_s \cdot ws.\text{foldl}((\cdot), I)\big),$$

which matches $M \cdot (s :: ws).\text{foldl}((\cdot), I)$ (`foldl_matrix_mul_assoc` in `HyperCubeGroup/SampleComplexity/CayleyIdentifiability.lean`). □

Lemma B.2 (Associativity of foldl)**.** *For an associative operation f, $w.\mathit{foldl}(f.\mathit{op}, a \circ b) = a \circ (w.\mathit{foldl}(f.\mathit{op}, b))$.*

Proof. By induction on w. For $w = []$ the claim is $a \circ b = a \circ b$. For $w = s :: ws$, associativity $f(f(a,b),s) = f(a, f(b,s))$ gives $ws.\text{foldl}(f.\text{op}, f(f(a,b),s)) = ws.\text{foldl}(f.\text{op}, f(a, f(b,s))) = a \circ (ws.\text{foldl}(f.\text{op}, f(b,s)))$, which equals $a \circ ((s :: ws).\text{foldl}(f.\text{op}, b))$ (`foldl_op_assoc`). □

Theorem B.3 (Deterministic Cayley identifiability via generating scaffold)**.** *Suppose that (1) $S \supseteq S_{\text{gen}}(e, \Gamma)$; (2) the parameter state $\widehat{\Theta}$ achieves exact local locking on S ($A_a B_b = C^\dagger_{a\circ b}$ for $(a,b) \in S$); and (3) $\widehat{\Theta}$ is gauge-normalized at the identity, $A_e = B_e = I$. (For unitary slices this is attainable without loss of generality via the gauge symmetry $\Theta \mapsto (UAV^\dagger, VBW^\dagger, WCU^\dagger)$ with $U = I$, $V = A_e$, $W = A_e B_e$; it does not change the scale of non-unitary factors.) Then:*

1. ***Gauge synchronization:*** *$A_g = B_g = C_g^\dagger$ for all $g \in Q$.*
2. ***Homomorphism on generators:*** *$A_g A_s = A_{g\circ s}$ for all $g \in Q$ and $s \in \Gamma$.*
3. ***Word induction to full table:*** *$A_a A_b = A_{a\circ b}$ for all $a, b \in Q$.*

4. ***Full table factorization:*** *if the character trace is orthogonal,* $\frac{1}{n}\operatorname{Tr}(A_x A_y^\dagger) = \delta_{xy}$*, then* $T_{abc}(\widehat{\Theta}) = \delta_{abc}$ *everywhere on* Q^3*.*

Proof. (1) Locking on the row $\{e\}\times Q$ gives $A_e B_b = C_b^\dagger$, hence $B_b = C_b^\dagger$; on the column $Q\times\{e\}$ it gives $A_a B_e = C_a^\dagger$, hence $A_a = C_a^\dagger$ (`gauge_synchronization_of_identityCross`). (2) Locking on the generator column $Q\times\{s\}$ gives $A_g B_s = C_{g\circ s}^\dagger$, i.e. $A_g A_s = A_{g\circ s}$ by (1). (3) For any $a, b \in Q$, since Γ generates G, there is a word $w \in \Gamma^*$ with $b = \text{wordEval}(e, w)$. Lemma B.1 with $M = A_a$ gives $A_a \cdot \text{wordMatrixProd}(A, w) = w.\text{foldl}((\cdot), A_a)$. By the generator steps, folding along w evaluates to $A(w.\text{foldl}(f.\text{op}, a))$, and Lemma B.2 with $b = e$ gives $w.\text{foldl}(f.\text{op}, a) = a \circ (w.\text{foldl}(f.\text{op}, e)) = a \circ b$. The same argument with $a = e$ gives $\text{wordMatrixProd}(A, w) = A_b$. Therefore $A_a A_b = A_{a\circ b}$ (`homomorphism_of_generating_scaffold`). (4) With (1) and (3), $T_{abc} = \frac{1}{n}\operatorname{Tr}(A_{a\circ b} A_c^\dagger) = \delta_{abc}$ by trace orthogonality. □

Theorem B.4 (Canonical column scaffold)**.** *Let* $G = (Q, \circ)$ *be a finite group of order* n *with identity* e*, let* $\Gamma \subseteq Q$ *generate* G*, let* $\lambda > 0$*, and let* $S \supseteq S_{\text{gen}}(e, \Gamma)$ *(Definition 5). Let* $\widehat{\Theta}$ *be an exact global minimizer of* $\mathcal{J}_{\lambda,\Omega_S}$*, normalize* $\widetilde{A} = s_\lambda^{-1}\widehat{A}$*,* $\widetilde{B} = s_\lambda^{-1}\widehat{B}$*,* $\widetilde{C} = s_\lambda^{-1}\widehat{C}$*, and apply a gauge transformation so that* $\widetilde{A}_e = \widetilde{B}_e = I_n$*. Then* $\widetilde{A}_g = \widetilde{B}_g = \widetilde{C}_g^\dagger =: U_g$ *for all* g*,* $U_a U_b = U_{a\circ b}$ *for all* $a, b \in Q$*, and* $T_{abc}(\widehat{\Theta}) = s_\lambda^3 \delta_{abc}$ *on* Q^3*. Moreover* $|S_{\text{gen}}(e, \Gamma)| \le (2n-1) + |\Gamma|(n-1)$*, which is* $3n-2$ *for* $|\Gamma| = 1$ *and at most* $4n-3$ *for* $|\Gamma| = 2$*.*

Proof. $S_{\text{cross}} \subseteq S$ has three-slot coverage, so by Theorem 4 the normalized slices $\widetilde{A} = s_\lambda^{-1}\widehat{A}$ (and likewise for B, C) are unitary and satisfy $\widetilde{A}_a \widetilde{B}_b \widetilde{C}_{a\circ b} = I$ on S. The gauge normalization of Theorem B.3 applies to these unitary normalized slices and leaves T unchanged; parts (1)–(3) give gauge synchronization and the representation law. For (iii), fix any $(a_0, b_0) \in S$ with $g_0 = a_0 \circ b_0$. By Theorem 4, $T_{a_0 b_0 c}(\widehat{\Theta}) = s_\lambda^3 \mathbf{1}_{\{c = g_0\}}$ for all c, i.e. $\frac{1}{n}\operatorname{Tr}(U_{g_0} U_c^\dagger) = \mathbf{1}_{\{c=g_0\}}$. Theorem 7 applied to U gives $\frac{1}{n}\operatorname{Tr}(U_a U_b U_c^\dagger) = \delta_{abc}$, so $T(\widehat{\Theta}) = s_\lambda^3 \delta$. The cardinality bound is $|S_{\text{cross}}| + |\Gamma|(n-1)$, since each generator column adds at most $n-1$ pairs outside the row $\{e\}\times Q$ (`card_generatingScaffold_le_idealized`, `card_generatingScaffold_cyclic_le`, `card_generatingScaffold_two_generators_le`). □

B.2 Minimal presentation scaffold rigidity

Theorem B.5 (Minimal presentation scaffold)**.** *Let* $G = (Q, \circ)$ *be a finite group of order* n *with identity* e*, let* $\Gamma \subseteq Q$ *generate* G*, and let* R *be a set of defining relators: words over* $\Gamma \cup \Gamma^{-1}$ *such that* $s \mapsto s$ *induces an isomorphism* $\langle \Gamma \mid R\rangle \to G$*. Let* T_{Cay} *be a directed spanning tree of* $\operatorname{Cay}(G, \Gamma)$ *rooted at* e*. A relator* $r = s_{i_1}^{\varepsilon_1} \cdots s_{i_k}^{\varepsilon_k}$ *is traced from* e *as the closed path* $(p_0, t_1), (p_1, t_2), \ldots, (p_{k-1}, t_k)$*, where* $t_j = s_{i_j}^{\varepsilon_j} \in G$ *is the element named by the* j*-th letter,* $p_0 = e$*,* $p_j = p_{j-1} \circ t_j$*, and* $p_k = e$*; let* R_{indep} *consist of the traced pairs outside* $S_{\text{cross}} \cup T_{\text{Cay}}$*, and* $S_{\min} = S_{\text{cross}} \cup T_{\text{Cay}} \cup R_{\text{indep}}$*. Assume* inverse-letter consistency*:* $(s^{-1}, s) \in S_{\min}$ *for every* $s \in \Gamma$ *whose inverse occurs as a letter of a relator. Let* $\lambda > 0$*,* $S \supseteq S_{\min}$*, and let* $\widehat{\Theta}$ *be an exact global minimizer of* $\mathcal{J}_{\lambda,\Omega_S}$*, normalized and gauge-fixed as in Theorem B.4. Then* $\widetilde{A}_g = \widetilde{B}_g = \widetilde{C}_g^\dagger =: U_g$ *for all* g*,* $U_a U_b = U_{a\circ b}$ *for all* $a, b \in Q$*, and* $T_{abc}(\widehat{\Theta}) = s_\lambda^3 \delta_{abc}$ *on* Q^3*. Moreover* $|S_{\min}| \le (2n-1) + (n-1) + |R_{\text{indep}}| = 3n - 2 + |R_{\text{indep}}|$*.*

For the S_5 relators of Appendix G, inverse-letter consistency holds because tracing σ^5 visits $(\sigma^4, \sigma) = (\sigma^{-1}, \sigma)$.

Proof. 1. **Metric and gauge locking.** As in the column case, Theorem 4 makes the normalized slices unitary and locked on S, and the gauge normalization at e synchronizes them into one family $U_g = \widetilde{A}_g = \widetilde{B}_g = \widetilde{C}_g^\dagger$ with $U_e = I$ and $U_a U_b = U_{a\circ b}$ for $(a, b) \in S$. Unitary gauge transformations act on the normalized slices; they cannot remove the common factor s_λ of the raw slices.

2. **Element naming via the spanning tree.** Each of the $n-1$ tree edges (g, s) gives $U_g U_s = U_{g\circ s}$. Induction along the unique tree path from e to g, with word w_g over Γ, gives $U_g = w_g(U_\gamma : \gamma \in \Gamma)$.

3. **Relator cycle closure.** For each relator, locking at the traced pairs gives $U_{p_{j-1}} U_{t_j} = U_{p_j}$; chaining from $U_{p_0} = I$ to $U_{p_k} = U_e = I$ yields $U_{t_1} \cdots U_{t_k} = I$. For an inverse letter $t_j = s^{-1}$, locking at (s^{-1}, s) gives $U_{s^{-1}} U_s = U_e = I$, so $U_{t_j} = U_s^{-1}$. Hence $U_{s_{i_1}}^{\varepsilon_1} \cdots U_{s_{i_k}}^{\varepsilon_k} = I$ for every relator, and von Dyck's theorem (Lyndon & Schupp, 1977) provides a homomorphism $\varphi : G \to \mathrm{U}(n)$ with $\varphi(s) = U_s$ for $s \in \Gamma$.

4. **Full table identifiability.** By step 2, $U_g = w_g(U_\Gamma) = \varphi(w_g) = \varphi(g)$ for every g, so $U = \varphi$ is a representation and $U_a U_b = U_{a\circ b}$ for all pairs, observed or not. The full-table factorization follows from the anchor argument in the proof of Theorem B.4.

The cardinality bound is $|S_{\min}| \le |S_{\text{cross}}| + |T_{\text{Cay}}| + |R_{\text{indep}}| = (2n-1) + (n-1) + |R_{\text{indep}}|$ (`card_minimalPresentationScaffold_le`). R_{indep} is the set of distinct traced pairs; it need not be a smallest set of pairs that identifies G. In the Lean development, steps 2 and 3 are replaced by the hypothesis `PresentationRelatorClosed`, which asserts that locking on $T_{\text{Cay}} \cup R_{\text{indep}}$ implies the generator steps $U_g U_s = U_{g\circ s}$ for all $g \in Q$ and $s \in \Gamma$; step 1 and the word induction from generator steps to the full representation law are mechanized (Appendix M). □

B.3 Proof of the character anchor theorem

Proof of Theorem 7. For any $h \in G$, define $c = h^{-1} g_0$. Because R is a unitary representation, $R_c^\dagger = R_{c^{-1}} = R_{g_0^{-1} h}$, and $R_{g_0} R_c^\dagger = R_{g_0} R_{g_0^{-1} h} = R_h$. Moreover, $c = g_0 \iff h^{-1} g_0 = g_0 \iff h = e$. Substituting into the single fiber condition yields $\frac{1}{n}\chi_R(h) = \frac{1}{n}\operatorname{Tr}(R_h) = \mathbf{1}_{\{h=e\}}$, which is the character of the left regular representation ρ_{reg} (`one_fiber_forces_regular_character`, `one_fiber_forces_character_eq_regular`). Two complex representations of a finite group are isomorphic if and only if they have identical characters (Serre, 1977); here, with $\langle \chi, \psi \rangle = \frac{1}{n}\sum_g \chi(g)\overline{\psi(g)}$, one has $\langle \chi_R, \chi_{\text{reg}} \rangle = \|\chi_{\text{reg}}\|^2 = n$ (`regular_character_norm_sq`). Because both representations are unitary, the intertwiner can be taken unitary: there exists $U \in \mathrm{U}(n)$ with $R_g = U \rho_{\text{reg}}(g) U^\dagger$ for all $g \in G$; this step (ii) is not formalized in Lean. Part (iii) uses only (i): for any triple,

$$T_R(a,b,c) = \frac{1}{n}\operatorname{Tr}(R_a R_b R_c^\dagger) = \frac{1}{n}\operatorname{Tr}(R_{ab} R_c^\dagger) = \frac{1}{n}\operatorname{Tr}(R_{c^{-1}ab}) = \mathbf{1}_{\{c^{-1}ab=e\}} = \delta_{abc},$$

so the complete Cayley tensor is recovered on G^3 (`one_fiber_forces_full_cayley_recovery`). □

Relation to the negative-target error. Under $\mathcal{J}_{\lambda,\Omega_S}$, every observed pair of a global minimizer carries zero negative error and positive value s_λ^3 (Theorem 4); after normalization, each observed pair is therefore an anchor. Proposition A.5 characterizes zero negative error at all observed pairs; Theorem 7 shows that one pair already forces the regular character.

B.4 Virtual anchor frames for group isotopes

For a general group isotope without an identity element, choose an arbitrary reference pair $(a_0, b_0) \in Q \times Q$ and define the row and column bijections $L_{a_0}(b) := a_0 \circ b$ and $R_{b_0}(a) := a \circ b_0$. The virtual anchor cross is $S_{\text{anchor}}(a_0, b_0) := (\{a_0\} \times Q) \cup (Q \times \{b_0\})$. For unitary (normalized) slices, the gauge symmetry $\Theta \mapsto (UAV^\dagger, VBW^\dagger, WCU^\dagger)$ with $U = I$, $V = A_{a_0}$, and $W = A_{a_0} B_{b_0}$ normalizes $A_{a_0} = I$, $B_{b_0} = I$, and $C_{a_0 \circ b_0} = I$. Local locking $A_a B_b = C_{a\circ b}^\dagger$ on S_{anchor} then forces $B_b = C_{L_{a_0}(b)}^\dagger$ and $A_a = C_{R_{b_0}(a)}^\dagger$. This collapses the gauge freedom $(U, V, W) \in \mathrm{U}(n)^3$ to a single representation frame, mapping the remaining generator relations to the underlying group presentation.

C OBSERVED-PAIR GRAPH CONNECTIVITY (BACKGROUND)

View an observation set $S \subseteq Q \times Q$ as a bipartite graph with row vertices V_A, column vertices V_B ($|V_A| = |V_B| = n$), and one edge per observed pair. A connected graph on $2n$ vertices has at least $2n-1$ edges, so a spanning tree is the smallest connected support. Connectivity, even together with three-slot coverage, does not synchronize the three factor families: the five-edge C_3 tree of Section 4.4 is connected and covering, yet admits an exact global minimizer with a residual sign ambiguity. The identity cross is a particular spanning tree—a star centred at the row and column of e—whose gauge normalization synchronizes the families (Proposition 8); it still does not determine unobserved products. Under random sampling, synchronization is instead supplied by the comparison graphs of Appendix L. For a Bernoulli-p mask, the expected number of isolated vertices of the input-pair graph is $2n(1-p)^n$, which is small once pn exceeds $\log(2n)$; the comparison-graph condition (7) requires $p^2 n$ of order $\log n$.

C.1 A SPANNING-TREE COUNTEREXAMPLE ON C_3

The identity cross is a spanning tree in the bipartite input-pair graph whose vertices are row and column labels and whose edges are observed pairs. An arbitrary spanning tree of the same size need not even synchronize the factor families.

Consider the cyclic group $C_3 = \mathbb{Z}_3$, with arithmetic modulo 3, and the five-edge spanning tree

$$T = \{(0,0), (0,1), (1,1), (1,2), (2,2)\}.$$

It has size $2n - 1 = 5$, connects all row vertices $\{0,1,2\}$ and column vertices $\{0,1,2\}$, and covers all output symbols, so it has three-slot coverage.

Let L_g denote the regular representation of C_3 and consider the unitary parameter state Θ_{bad} obtained by scalar sign twists,

$$A_a = \alpha_a L_a, \qquad B_b = \beta_b L_b, \qquad C_c = \gamma_c L_c^\dagger,$$
$$\alpha = (1,-1,-1), \qquad \beta = \gamma = (1,1,-1).$$

Because L_g orthogonalizes incorrect targets,

$$T_{abc}(\Theta_{\text{bad}}) = \alpha_a \beta_b \gamma_c \, \mathbf{1}_{\{a+b=c\}},$$

so every negative candidate vanishes. On the five observed pairs,

$(a,b) \mapsto a+b$	$(0,0) \mapsto 0$	$(0,1) \mapsto 1$	$(1,1) \mapsto 2$	$(1,2) \mapsto 0$	$(2,2) \mapsto 1$
$\alpha_a \beta_b \gamma_{a+b}$	$1 \cdot 1 \cdot 1$	$1 \cdot 1 \cdot 1$	$(-1) \cdot 1 \cdot (-1)$	$(-1) \cdot (-1) \cdot 1$	$(-1) \cdot (-1) \cdot 1$

and every product equals 1.

Hence Θ_{bad} achieves exact categorical fitting (score 1 on valid targets and 0 on invalid targets) on all five observed tree edges. For $\lambda > 0$, $s_\lambda \Theta_{\text{bad}}$ attains the lower bound in Theorem 4 and is therefore an exact global minimizer of $\mathcal{J}_{\lambda,\Omega_T}$. Yet on the unobserved pair $(1,0)$ with valid output $1 + 0 = 1$,

$$T_{1,0,1}(\Theta_{\text{bad}}) = \alpha_1 \beta_0 \gamma_1 = (-1) \cdot 1 \cdot 1 = -1 \neq 1.$$

The two incorrect-output scores are zero, so the output argmax is incorrect at $(1,0)$, also after scaling by $s_\lambda > 0$.

Synchronization also fails. The normalized tuple already has $A_0 = B_0 = C_0^\dagger = I_3$, but $A_1 = -B_1$. If a unitary gauge synchronized the three families into one family R_g, the observed law at $(0,0)$ would give $R_0^2 = R_0$, hence $R_0 = I_3$ by unitarity. Such a gauge must therefore preserve the identity normalization. The only unitary gauges preserving $A_0 = B_0 = I_3$ are simultaneous conjugations, which preserve the mismatch $A_1 = -B_1$. Thus no unitary gauge synchronizes the three families.

The tree thus achieves exact categorical fitting and metric locking on the observed support, yet leaves a residual phase ambiguity that corrupts unobserved entries, and Θ_{bad} does not factorize C_3 (formally verified; Appendix M). The normalized tuple satisfies full-output constraints at every observed pair, so output anchors are present; the representation-law hypothesis of Theorem 7 is missing. The failure is a phase ambiguity rather than a lack of coverage or connectivity: recovery needs further algebraic relations among observed products.

The counterexample leaves a specific question: which relations in a random mask eliminate the remaining phase ambiguity? The comparison graphs of Section 4.5 and Appendix L encode equalities obtained from pairs of observed products and propagate them into a common representation.

D EXTENSION TO GROUP ISOTOPES

A group isotope is a quasigroup $(Q, \circ)$ whose Cayley table is equivalent to a group table up to independent permutations of its rows, columns, and symbols: there exist a group $(Q, \cdot)$ and bijections $\phi, \psi, \chi \in S_n$ with $a \circ b = \chi(\phi^{-1}(a) \cdot \psi^{-1}(b))$.

Relabeling versus gauge. Because the three-factor architecture assigns independent matrix families to the row (A_a), column (B_b), and output (C_c) slots, it absorbs these coordinate permutations by relabeling slice indices: $\Theta \mapsto \Theta'$ with $A'_{\phi(a)} = A_a$, $B'_{\psi(b)} = B_b$, $C'_{\chi(c)} = C_c$ maps the group objective on S to the isotope objective on $\{(\phi(a), \psi(b)) : (a, b) \in S\}$, term by term, and preserves every local sensitivity. This relabeling is distinct from the matrix gauge symmetry, which fixes all labels and leaves the labeled prediction tensor unchanged. Statements that refer to no particular label—for instance Theorems 3, 4, 9 and A.7—therefore transfer to isotopes; statements that use the identity element transfer with the identity cross replaced by a virtual anchor cross (Appendix B.4).

D.1 UNCONDITIONAL METRIC AND SCALE LOCKING (TIER 1)

Theorem D.1 (Isotope metric locking)**.** *For any finite group isotope with three-slot coverage, every exact global minimizer of $\mathcal{J}_{\lambda,\Omega_S}$ has uniform scaled-unitary Gram matrices ($MM^\dagger = s_\lambda^2 I$) and full rank.*

Proof. By Lemma 14 of Huh et al. (2025), any quasigroup isotopic to a finite group admits an exact unitary collinear factorization Θ_{opt}. Uniformly scaling this unitary factorization by the equilibrium scale s_λ yields slices with $\widehat{A}_a\widehat{A}_a^\dagger = \widehat{B}_b\widehat{B}_b^\dagger = \widehat{C}_c\widehat{C}_c^\dagger = s_\lambda^2 I$, which attain the scalar envelope minimum $\phi_\lambda(s_\lambda)$ on every observed edge. Therefore, by the global minimization argument of Theorem 4, every exact global minimizer on a group isotope has this uniformly scaled, full-rank form. (Equivalently, apply Theorem 4 to the underlying group and relabel.) □

D.2 CAYLEY IDENTIFIABILITY VIA VIRTUAL ANCHOR FRAMES (TIER 2)

Unlike groups, a general group isotope need not possess a two-sided identity element e, which prevents the direct normalization of an identity cross. However, the three-factor model has the gauge symmetry group $\mathcal{G} = \mathrm{U}(n) \times \mathrm{U}(n) \times \mathrm{U}(n)$, acting as $\Theta \mapsto (UAV^\dagger, VBW^\dagger, WCU^\dagger)$, which allows us to construct a *virtual* identity frame. Fix any reference pair $(a_0, b_0) \in Q \times Q$, define $L_{a_0}(b) := a_0 \circ b$ and $R_{b_0}(a) := a \circ b_0$, and let $S_{\text{anchor}}(a_0, b_0) := (\{a_0\} \times Q) \cup (Q \times \{b_0\})$. For the normalized unitary slices, choosing $U = I$, $V = A_{a_0}$, and $W = A_{a_0}B_{b_0}$ gives $A_{a_0} = I$, $B_{b_0} = I$, and $C_{a_0 \circ b_0} = I$. Since Theorem D.1 and the equality case of Theorem 4 give the local locking condition $A_aB_b = C_{a\circ b}^\dagger$ on all observed edges, applying it along the two arms of S_{anchor} forces $B_b = C_{L_{a_0}(b)}^\dagger$ and $A_a = C_{R_{b_0}(a)}^\dagger$. This synchronization collapses the $\mathrm{U}(n)^3$ gauge freedom to a single shared representation frame (equivalent to the $\mathrm{U}(n)$ diagonal conjugation of the shared-embedding model). The remaining generator observations then trace the underlying group presentation, identifying the complete Cayley table of the isotope up to the coordinate permutations (ϕ, ψ, χ).

E INFORMATION-THEORETIC COMPARISON VIA FANO'S INEQUALITY

The sample complexity hierarchy reflects an information-theoretic distinction between arbitrary quasigroups and group isotopes. Logarithms in this appendix are base 2.

Quasigroups. The number of Latin squares of order n scales as $L_n = \left((1 + o(1))\frac{n}{e^2}\right)^{n^2}$ (van Lint & Wilson, 2001). Specifying an arbitrary quasigroup without structural prior knowledge therefore requires $\log L_n = \Theta(n^2 \log n)$ bits, and because each observed entry $(a, b, a \circ b)$ at a target-independent location supplies at most $\log n$ bits, recovering an arbitrary quasigroup requires $\Omega(n^2)$ such observations.

Isotopes of a fixed group. Fix a group $G = (Q, \cdot)$ of order n. Its isotopes are the tables $a \circ b = \chi(\phi^{-1}(a) \cdot \psi^{-1}(b))$ with $(\phi, \psi, \chi) \in (S_n)^3$; two triples give the same table exactly when they differ by an autotopism of G, i.e. a triple (α, β, γ) of bijections with $\alpha(x)\beta(y) = \gamma(xy)$ for all x, y.

Lemma E.1 (Autotopisms of a group)**.** *The autotopism group of G has order $n^2|\mathrm{Aut}(G)|$. Hence G has exactly $N_G = (n!)^3/(n^2|\mathrm{Aut}(G)|)$ distinct isotope tables.*

Proof. Let (α, β, γ) be an autotopism and put $u = \alpha(e)$, $v = \beta(e)$. Setting $y = e$ and $x = e$ gives $\alpha(x) = \gamma(x)v^{-1}$ and $\beta(y) = u^{-1}\gamma(y)$, so $\gamma(x)\, w\, \gamma(y) = \gamma(xy)$ with $w = v^{-1}u^{-1}$. Then $\theta(x) := w\gamma(x)$ satisfies $\theta(x)\theta(y) = \theta(xy)$ and is bijective, so $\theta \in \mathrm{Aut}(G)$, and $(\alpha, \beta, \gamma) = (uv\,\theta(\cdot)\,v^{-1},\ v\,\theta(\cdot),\ uv\,\theta(\cdot))$. Conversely every $(u, v, \theta) \in G \times G \times \mathrm{Aut}(G)$ defines an autotopism by this formula, and (u, v, θ) is recovered as $(\alpha(e), \beta(e), w\gamma)$. The second claim is the orbit–stabilizer theorem. □

Since G is generated by at most $\log n$ elements and an automorphism is determined by the images of a generating set, $|\mathrm{Aut}(G)| \le n^{\log n}$. With $\log n! \ge n\log(n/e)$,

$$3n\log n \ \ge\ \log N_G \ \ge\ 3n\log n - 3n\log e - 2\log n - \log^2 n = 3n\log n - \mathcal{O}(n).$$

The class of all isotopes of all groups of order n is larger by at most a factor $n^{(2/27+o(1))\log^2 n}$, the number of isomorphism types (Pyber, 1993), which does not change the leading term.

Fano lower bound for target-independent observations. Let the target f^* be uniform over the N_G isotope tables of G, and let a learner query m cells whose locations may depend on earlier answers and on independent randomness, but not otherwise on f^*. Each answer is one of n labels, so the mutual information between f^* and the answers is at most $m \log n$, and Fano's inequality (Cover & Thomas, 2006) gives, for any estimator $\widehat{f}$ with $\Pr(\widehat{f} \ne f^*) \le \delta$,

$$m \ge \frac{(1-\delta)\log N_G - 1}{\log n} \ge (1-\delta)\left(3n - \mathcal{O}\left(\frac{n}{\log n}\right)\right).$$

For fixed δ the leading coefficient is $3(1-\delta)$. If only the labels of a group are unknown ($\phi = \psi = \chi$), the class has $n!/|\mathrm{Aut}(G)|$ tables and the same argument gives $(1-\delta)(n - \mathcal{O}(n/\log n))$.

Teacher-chosen locations. The bound above does not apply when a teacher chooses the observed set with knowledge of the target, because the choice of locations itself carries information. Counting labeled sets gives the general constraint for that model: distinct tables need distinct labeled observation sets, and there are at most $\sum_{k \le m} \binom{n^2}{k} n^k \le (en^3/m)^m$ labeled sets of size at most $m \le n^2$. Hence $m\log(en^3/m) \ge \log N_G$, which yields only $m \ge (3/2 - o(1))n$ for the isotope class and $m \ge (1/2 - o(1))n$ for relabeled copies of G.

Comparison with the active construction. Theorem B.5 recovers the Cayley table from at most $3n - 2 + |R_{\mathrm{indep}}|$ teacher-chosen pairs; because the Cayley spanning tree rooted at e contains at least one generator edge $(e, g) \in S_{\mathrm{cross}}$, the practical cardinality is at most $3n - 3 + |R_{\mathrm{indep}}|$ (Appendix G). When $|R_{\mathrm{indep}}| = o(n)$, its leading coefficient 3 equals the limit of the Fano coefficient $3(1-\delta)$ as $\delta \to 0$. The two bounds refer to different query models, so this agreement is a comparison of leading terms, not a minimax optimality statement for either model. We also do not infer $|R_{\mathrm{indep}}| = o(n)$ for general groups from results on short presentations: Babai et al. (1997) prove, using the classification of finite simple groups, that every finite group without composition factors ${}^2A_2(q) = \mathrm{PSU}(3, q)$, ${}^2B_2(q) = \mathrm{Sz}(q)$, and ${}^2G_2(q)$ has a presentation of length $\mathcal{O}(\log^3 n)$, but their length counts exponents in binary, whereas tracing a relator queries one pair per letter of the expanded word (for example, $x^n = 1$ has length $\mathcal{O}(\log n)$ and trace length n). For the benchmark presentations, $|R_{\mathrm{indep}}|$ is constant for the cyclic, dihedral, and fixed permutation groups and equals $k^2 = \log^2 n$ for $(C_2)^k$ (Table 1).

G Exact scaffold constructions and relator cycle traces

To connect Theorem B.5 with the experiments of Section 5, we detail the pair counts of the minimal presentation scaffolds $S_{\min}$ produced by the accompanying `dataset.py` module.

As in Definition 5, the minimal presentation scaffold is the union of the identity cross S_{cross}, a Cayley spanning tree T_{Cay}, and the relator pairs R_{indep}. Because a relator word corresponds to a closed loop in the Cayley graph, tracing it requires evaluating a sequence of intermediate binary operations; $|R_{\text{indep}}|$ denotes the number of distinct traced pairs (a, b) that are not already in $S_{\text{cross}} \cup T_{\text{Cay}}$. The overlap between relator cycles and the spanning tree depends on the chosen tree and on the prefixes of the relator words, so $|R_{\text{indep}}|$ has no universal closed form; it is evaluated as $|S_{\min}| - |S_{\text{cross}} \cup T_{\text{Cay}}|$.

Deduplication. Because T_{Cay} rooted at e contains at least one generator edge leaving the identity, $(e, g) \in S_{\text{cross}}$, one has $|S_{\text{cross}} \cap T_{\text{Cay}}| \geq 1$, so the practical cardinality is at most $3n - 3 + |R_{\text{indep}}|$, within the mechanized bound $3n - 2 + |R_{\text{indep}}|$. For the trees built by the code, all $|\Gamma|$ edges (e, s), $s \in \Gamma$, are tree edges, so $|S_{\text{cross}} \cup T_{\text{Cay}}| = 3n - 2 - |\Gamma|$.

Multi-generator compression. The canonical column scaffold S_{gen} suggests a sample complexity of $(|\Gamma|+2)n-(|\Gamma|+1)$ that grows with the number of generators. The minimal presentation scaffold removes this redundancy: to reach and name all n elements the learner needs a directed Cayley spanning tree ($n - 1$ edges), and to identify the presentation among non-isomorphic candidates it needs the $|R_{\text{indep}}|$ relator pairs (with the code's trees, $|R_{\text{indep}}| = 4$ for V_4 and D_8, and 6 for Q_8; Table 1). By replacing full generator columns with this skeleton, the active recovery threshold no longer scales with $|\Gamma|$ beyond $|R_{\text{indep}}|$.

G.1 Algorithmic construction of $S_{\min}$ and $|R_{\text{indep}}|$ in `dataset.py`

The construction is implemented by `get_group_scaffolds` in `dataset.py`.

1. **Foundational base (`tree_base`).** The identity cross is built by `get_identity_cross` ($2n - 1$ pairs). The Cayley spanning tree is built by `build_cayley_spanning_tree`: starting from the identity, it visits nodes from a stack and, for each generator, follows that generator's powers from the current node, recording the pair `(curr, gen)` whenever an unvisited element is reached ($n - 1$ pairs). The two lists are merged into the set `tree_base=set(cross+tree_edges)`; because the cross and the tree share the initial generator edges `(e, g)`, the union has at most $3n - 3$ elements. For cyclic groups the code uses the generator column itself, which is a spanning path of the Cayley graph.
2. **Cycle tracing (`trace_cycle`).** For each relator word w in `GROUP_PRESENTATION_SPECS`, `trace_cycle` steps through the Cayley table from the identity (`g0=0`), logging each queried pair `(curr, g)` and updating the state to the product. An inverse letter is stored as the index of the inverse element (σ^{-1} for S_5) or rewritten as a positive power ($j^{-1} = j^3$ for Q_8).
3. **Set-union deduplication (`s_min_pairs`).** The relator pairs are added by `tree_base.union(relator_edges)`, so pairs already in the cross or the tree are absorbed.
4. **Isolating $|R_{\text{indep}}|$.** $|R_{\text{indep}}| = |S_{\min}| - |S_{\text{cross}} \cup T_{\text{Cay}}|$.

Table 1: Presentation scaffolds produced by the supplied `dataset.py`. For every row, the group presented by the relations $w_a w_b = w_{ab}$, $(a, b) \in S_{\min}$ (w_g the tree word of g), has order exactly n (SymPy coset enumeration; checked for every row with a numerical n and for D_8, D_{24}, $(C_2)^3$, $(C_2)^7$).

Group	n	Relators	$\lvert S_{\text{cross}} \cup T_{\text{Cay}}\rvert$	$\lvert R_{\text{indep}}\rvert$	$\lvert S_{\min}\rvert$	$\lvert S_{\min}\rvert - 3n$
C_n	n	g^n	$3n-3$	1	$3n-2$	-2
D_n	$n=2k$	$s^k,\ r^2,\ (sr)^2$	$3n-4$	4	$3n$	0
$(C_2)^k$	2^k	$g_i^2,\ (g_i g_j)^2$	$3n-2-k$	k^2	$3n-2-k+k^2$	k^2-k-2
V_4	4	$a^2,\ b^2,\ (ab)^2$	8	4	12	0
Q_8	8	$i^4,\ j^4,\ i^2j^2,\ ijij^3$	20	6	26	2
S_3	6	$\tau^2,\ \sigma^3,\ (\tau\sigma)^2$	14	4	18	0
A_4	12	$\sigma^3,\ \tau^3,\ (\sigma\tau)^2$	32	4	36	0
S_4	24	$\tau^2,\ \sigma^4,\ (\tau\sigma)^3$	68	6	74	2
A_5	60	$\sigma^3,\ \tau^5,\ (\sigma\tau)^5,\ (\sigma\tau^2)^2$	176	8	184	4
S_5	120	see text	356	19	375	15

G.2 GROUP FAMILY SCAFFOLD CASE STUDIES

Cyclic groups C_n. A single generator g with relator $g^n = e$. The spanning tree is the path of length $n-1$ along column g; the cross and tree share (e, g), giving a base of $(2n-1)+(n-1)-1 = 3n-3$. Closing $g^n = e$ requires the off-tree pair (g^{n-1}, g), so $|R_{\text{indep}}| = 1$ and $|S_{\min}|(n) = 3n-2$, which coincides with the column scaffold.

Elementary Abelian 2-groups $(C_2)^k$, $n = 2^k$. Generators are k independent bit masks $1, 2, 4, \ldots, 2^{k-1}$, with relators $g_i^2 = e$ (k relations) and $(g_i g_j)^2 = e$ for $i < j$ ($\binom{k}{2}$ relations). The base set shares k generator edges leaving the identity, of size $3n - 2 - k$. With the code's tree, each involution relator adds one pair and each commutation relator adds two, so $|R_{\text{indep}}| = k + 2\binom{k}{2} = k^2 = \log_2^2 n$ and $|S_{\min}| = 3n - 2 - k + k^2$ (28 for $(C_2)^3$, 424 for $(C_2)^7$). A tree that contains the edge (g_i, g_j) for every $i < j$ (for example, a breadth-first tree over the bit masks) saves one pair per commutation relator and gives $|R_{\text{indep}}| = k(k+1)/2$, the count stated in the working draft (11 for $V_4 \cong (C_2)^2$).

Quaternion group Q_8. Generators i, j; relators i^4, j^4, $i^2 j^2$, $iji j^{-1}$ (with j^{-1} written as j^3). The cross (15 pairs) and tree (7 edges) share (e, i) and (e, j), giving 20 pairs; tracing the four relators adds 6 pairs, so $|S_{\min}|(8) = 26 = 3n + 2$. The 24-pair Q_8 scaffold of Appendix J is a different construction (one generator column plus two relator pairs).

Dihedral groups D_{2k}, $n = 2k$. A rotation s of order k and a reflection r of order 2, with relators $s^k = 1$, $r^2 = 1$, $(sr)^2 = 1$. The base has $3n-4$ pairs, and tracing the three relators adds $1+1+2 = 4$ pairs, so $|S_{\min}|(n) = 3n$ for every dihedral group of the benchmark (24 for D_8).

Symmetric groups S_k. Two-generator presentations of S_k (such as the Moore presentation) use relator words whose lengths and number change with k, and the overlap of the relator loops with the chosen spanning tree varies, so no single elementary closed form is claimed. Structurally, $|S_{\min}|(k) = 3(k!)-4+|R_{\text{indep}}(k)|$, with $|R_{\text{indep}}(k)|$ evaluated by set union. For S_5 ($n = 120$, table size 14,400), with $\tau = (0\ 1)$, $\sigma = (0\ 1\ 2\ 3\ 4)$, and relators τ^2, σ^5, $(\tau\sigma)^4$, $(\tau\sigma^2\tau\sigma^3)^2$, $(\tau\sigma\tau\sigma^{-1})^3$: the cross has 239 pairs and the tree 119 edges, sharing 2, for a base of 356 pairs; tracing the five relators adds $1 + 1 + 6 + 4 + 7 = 19$ pairs, so the current constructor gives $|S_{\min}| = 375$ (2.60% of the table). Tracing σ^5 visits (σ^{-1}, σ), so inverse-letter consistency (Appendix B.2) holds. The S_5 experiment of Table 5 used a 371-pair scaffold (base 356 plus 15 relator pairs) produced by an earlier version of the constructor, whose pair list is not part of the supplied materials; Appendix J.5 reports runs on the 375-pair scaffold.

Worked fractions. For S_4 ($n = 24$, 576 entries, two generators), the column bound gives $m^+ \leq 4 \cdot 24 - 3 = 93$ observations (16.1%).

the presentation scaffold has 74 pairs (12.8%). For S_5, the column bound gives $m^+ \leq 4 \cdot 120 - 3 = 477$ entries (3.31%), and the presentation scaffold gives 375 entries (2.60%; 371 in the historical run).

I EMPIRICAL SCALING DATA ACROSS 32 BENCHMARK GROUPS

Benchmark groups. The 32 groups span orders $n \in [4, 128]$ in four families: (1) cyclic C_n (1 generator), $n \in \{4, 6, 8, 12, 16, 24, 36, 48, 60, 90, 120\}$ (11 groups); (2) dihedral D_n of order n (2 generators), $n \in \{4, 6, 8, 12, 16, 24, 36, 48, 60, 90, 120\}$ (11 groups); (3) permutation groups: symmetric S_3, S_4, S_5 ($n = 6, 24, 120$) and alternating A_4, A_5 ($n = 12, 60$) (5 groups); (4) elementary Abelian $(C_2)^3, \ldots, (C_2)^7$ ($n \in \{8, 16, 32, 64, 128\}$; k generators) (5 groups). The 32 tables represent 31 isomorphism types: D_6 and S_3 are isomorphic and enter with different labelings and presentations.

Thresholds. Table 2 lists $m^*(95\%)$ and $m^*(100\%)$ from `passive_sampling_results.json` under the criteria of Section 5. Where no tested m met the 100% criterion, the entry reads "n.r." (not reached within the tested grid); the analysis script otherwise substitutes the largest tested m, and those substituted values are not reported. All entries were recomputed from the per-run accuracies stored in the same file; every $m^*(95\%)$ is attained within the grid, and all 32 thresholds satisfy $0.91 \leq m^*(95\%)/m_{\text{hybrid}}(n) \leq 1.18$.

Table 2: Empirical sample complexity thresholds across 32 groups ($n = 4$ to 128).

Group	Family	n	n^2	$m^*(95\%)$	$p^*(95\%)$	$m^*(100\%)$	$p^*(100\%)$
C_4	Cyclic	4	16	11	68.8%	11	68.8%
D_4	Dihedral	4	16	12	75.0%	12	75.0%
C_6	Cyclic	6	36	20	55.6%	20	55.6%
D_6	Dihedral	6	36	18	50.0%	20	55.6%
S_3	Permutation	6	36	20	55.6%	20	55.6%
C_8	Cyclic	8	64	30	46.9%	n.r. (4/5 at 30)	–
D_8	Dihedral	8	64	27	42.2%	27	42.2%
$(C_2)^3$	Elem. Abelian	8	64	30	46.9%	30	46.9%
C_{12}	Cyclic	12	144	41	28.5%	48	33.3%
D_{12}	Dihedral	12	144	43	29.9%	48	33.3%
A_4	Permutation	12	144	41	28.5%	n.r. (4/5 at 45)	–
C_{16}	Cyclic	16	256	61	23.8%	64	25.0%
D_{16}	Dihedral	16	256	57	22.3%	64	25.0%
$(C_2)^4$	Elem. Abelian	16	256	64	25.0%	67	26.2%
C_{24}	Cyclic	24	576	104	18.1%	117	20.3%
D_{24}	Dihedral	24	576	98	17.0%	n.r. (4/5 at 123)	–
S_4	Permutation	24	576	98	17.0%	111	19.3%
$(C_2)^5$	Elem. Abelian	32	1,024	159	15.5%	187	18.3%
C_{36}	Cyclic	36	1,296	188	14.5%	188	14.5%
D_{36}	Dihedral	36	1,296	188	14.5%	188	14.5%
C_{48}	Cyclic	48	2,304	299	13.0%	299	13.0%
D_{48}	Dihedral	48	2,304	282	12.2%	282	12.2%
C_{60}	Cyclic	60	3,600	431	12.0%	431	12.0%
D_{60}	Dihedral	60	3,600	409	11.4%	409	11.4%
A_5	Permutation	60	3,600	431	12.0%	431	12.0%
$(C_2)^6$	Elem. Abelian	64	4,096	472	11.5%	472	11.5%
C_{90}	Cyclic	90	8,100	835	10.3%	835	10.3%
D_{90}	Dihedral	90	8,100	795	9.8%	795	9.8%
C_{120}	Cyclic	120	14,400	1,414	9.8%	1,414	9.8%
D_{120}	Dihedral	120	14,400	1,355	9.4%	1,355	9.4%
S_5	Permutation	120	14,400	1,414	9.8%	1,414	9.8%
$(C_2)^7$	Elem. Abelian	128	16,384	1,543	9.4%	1,543	9.4%

Table 3: Empirical scaling summary. Slopes are from regressions of $\ln m^*(95\%)$ on $\ln n$; uncertainties are ± 1.96 standard errors. Observation fractions are $m^*(95\%)/n^2$.

Quantity	Result
Hybrid fit	$R^2 = 0.9980$, RMSE $= 20.92$
Slope, all 32 groups	1.401 ± 0.044
Slope, 18 groups with $n \geq 24$	1.633 ± 0.031
Observation fraction, C_4	68.8%
Observation fraction, S_5 ($n = 120$)	9.8%
Observation fraction, $(C_2)^7$ ($n = 128$)	9.4%

Secondary plateau statistics. The normalized bridge ratio $c_{\text{bridge}}(n) = m^*(95\%)/(n^{1.5}\sqrt{\ln n})$ has mean 0.464 ± 0.018 (coefficient of variation 3.9%) over the 18 groups with $n \geq 24$; for the 11 groups with $n \leq 12$, coupon collection raises the ratio to a mean of 0.906.

Statistical assumptions. The hybrid coefficient is chosen by a grid search over $c \in [0.20, 1.00]$ (step 10^{-3}) minimizing the squared error of $\max(3n, n\ln(3n), c\, n^{1.5}\sqrt{\ln n})$ against the 32 values of $m^*(95\%)$, with R^2 computed against the mean. The power law is an ordinary least-squares fit of $\ln m^*(95\%)$ on $\ln n$; the reported interval is $\alpha \pm 1.96\,\text{SE}$, with $\text{SE} = 0.0226$ from the residual variance on 30 degrees of freedom, and the t-statistic is $(\alpha - 2)/\text{SE}$. The intermediate exponent reflects the composite fit: below the crossover the coupon floor $n\ln(3n)$ has local log-slope $1 + 1/\ln(3n) \in [1.2, 1.4]$, and above it the term $n^{1.5}\sqrt{\ln n}$ has local log-slope $1.5 + 1/(2\ln n) \in$

$[1.603, 1.657]$ on $24 \le n \le 128$. Over the 18 groups with $n \ge 24$, the power-law fit gives $\alpha = 1.633$ with SE $= 0.0157$ on 16 degrees of freedom, i.e. the interval $[1.602, 1.663]$. The same grid search with third term $c\, n^{5/3}(\ln n)^{1/3}$ selects $c = 0.282$ with $R^2 = 0.9990$ and RMSE 14.46, so R^2 on the linear scale does not discriminate between the two orders. Comparisons of local log-slopes do not identify logarithmic factors on this range either: the expression of Theorem 9 itself, $n^{3/2}(\log(8n^2/\delta))^{1/2}$ at $\delta = 0.05$, has local log-slope 1.568–1.587, outside the interval, whereas $n^{3/2}\ln(20n)$ (1.627–1.662) and $n^{5/3}(\ln n)^{-1/5}$ (1.604–1.625) lie inside it, and on $24 \le n \le 128$ the factor $n^{1/6}$ changes by 1.32, as much as $(\ln n)^{0.66}$. The empirical scaling is therefore reported as consistent with growth of order $n^{3/2}$ up to logarithmic factors, not as an estimate of the asymptotic rate. These computations treat the 32 thresholds as independent observations with homoscedastic errors in log space; each threshold is quantized to its group's grid of tested m values. The fit covers a finite range of orders and does not by itself establish an asymptotic rate.

Sufficient condition versus measured thresholds. At $\delta = 0.05$, the explicit sufficient condition (7) requires $p \ge (6\log(8n^2/\delta)/n)^{1/2}$. The right-hand side exceeds 1 for $n \le 83$ and is approximately 0.97, 0.86, and 0.83 at $n = 90$, 120, and 128, respectively. The measured thresholds $p^*(95\%)$ at these orders are 9.8%–10.3%, 9.4%–9.8%, and 9.4%, respectively (Table 2).

J Diagnostic case studies

This appendix records the small-group scaffold experiments ($C_4, V_4, C_6, C_8, Q_8, D_8$), the S_4 regularization sweep, and the active-teaching experiments on S_5 (historical 371-pair and current 375-pair scaffolds). Each value below is read from the named result file; settings are those of the producing script in the supplied code (Adam, sum reduction, h-normalization) unless stated.

J.1 Identity cross ambiguity (C_4 versus V_4)

On the identity cross S_{cross} ($2n - 1 = 7$ pairs for $n = 4$), $\mathbb{Z}_4$ and the Klein four-group V_4 have identical observations since $x \circ e = e \circ x = x$ for all x. In `results/c4_v4_results.json` ($\lambda \in \{0.03, 0.1, 0.3\}$, learning rate 0.025, 600 epochs, 10 restarts each), unobserved accuracy on S_{cross} is 27.8%–44.4% for C_4 and 30.0%–33.3% for V_4 (chance 25%), with no perfect run for either group at any λ. Because the $(n-1)^2 = 9$ unobserved entries contain no generator column or relator constraint, the objective cannot distinguish non-isomorphic groups sharing the identity cross (Proposition 8). Metric locking on S_{cross} is visible in the larger groups: for the quaternion group Q_8 ($n = 8$; `results/relator_cycle_closure_results.json`, $\lambda = 0.5$, 5 runs), S_{cross} gives 16.7% unobserved accuracy (chance 12.5%) with mean slice condition number 1.000005 and Gram deviation 3.2×10^{-6}, and for C_6 and D_8 the mean condition number on S_{cross} is 1.0 (`results/scaffold_sharpness_results.json`).

J.2 Generating column sharpness (C_4, C_6, C_8)

For C_4, the complete generator column ($3n - 2 = 10$ pairs) gives mean unobserved accuracy 93.3%, 98.3%, and 98.3% (8/10, 9/10, and 9/10 perfect runs) at $\lambda = 0.03$, 0.1, and 0.3 (`results/c4_v4_results.json`). Along the sequence of scaffolds from S_{cross} to S_{gen} stored in `results/c4_v4_curve_results.json`, the 9-pair set ($3n - 3$) gives 81.4% (6/10 perfect) and the 10-pair column gives 100.0% (10/10). For C_6 ($n = 6$; `results/scaffold_sharpness_results.json`, $\lambda = 0.5$, learning rate 0.025, 1000 epochs, 5 runs), removing one or two generator edges (15 or 14 pairs) gives 60.0% (0/5) or 44.5% (1/5) unobserved accuracy, whereas the complete column ($3n - 2 = 16$ pairs) reaches 100.0% in all runs (5/5). Figure 3a shows the same pattern for C_8: the identity cross (15 pairs) and the cross with the generator column missing two or one edges (20 or 21 pairs) stay below 62% mean unobserved accuracy, and the complete column ($3n - 2 = 22$ pairs) reaches 94.3%.

J.3 Multi-generator relator closure (V_4, Q_8, D_8)

For groups with $|\Gamma| \ge 2$, observing one generator column ($3n - 2$ pairs) does not identify relations outside the generated cyclic subgroup. The Q_8 and D_8 values are from `results/relator_`

`cycle_closure_results.json` ($\lambda = 0.5$, learning rate 0.025, 1000 epochs, 5 runs); the scaffolds there extend the one-column scaffold by single relator pairs and differ from the $S_{\min}$ of Table 1.

- V_4 (`results/c4_v4_curve_results.json`): the one-column scaffold (10 pairs) identifies only $\langle a \rangle$, with 63.3% unobserved accuracy (5/10 perfect). Adding the relators $b^2 = e$ and $ab = ba$ gives $m = 12 = 3n$ pairs and 100.0% unobserved accuracy in all 10 restarts.
- Q_8: the one-column scaffold (22 pairs) identifies only $\langle i \rangle$ (47.6%, 0/5). Adding $j^2 = -1$ alone gives 70.7% and adding $ij = k$ alone 61.0% (both 0/5); adding both gives $m = 24 = 3n$ pairs and 100.0% unobserved accuracy (5/5).
- D_8: the one-column scaffold (22 pairs) gives 52.9% (0/5). Adding $r^2 = 1$ gives 61.5% (0/5) and adding $sr = rs^3$ gives 84.4% (3/5); adding both ($m = 24 = 3n$) gives 98.5% (4/5), and also observing $s^2r = rs^2$ ($m = 25 = 3n + 1$) gives 100.0% (5/5). At $\lambda = 10$ (`results/scaffold_sharpness_lam10_results.json`), the 24-pair presentation scaffold gives 100.0% in 5/5 runs, with per-run condition numbers between 1.00 and 1.68, while its 23-pair ablation without the $(sr)^2$ trace gives 55.1%.

Figure 3b plots a D_8 sequence of the same kind: identity cross, one generator column, then added relator pairs.

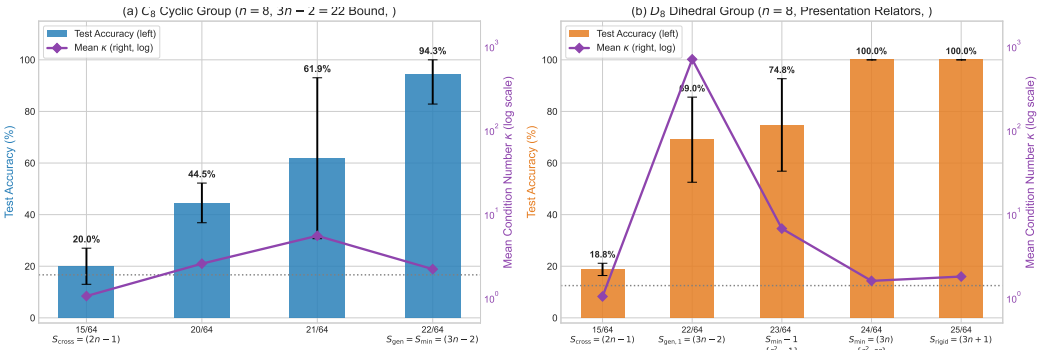


Figure 3: Scaffold ablations for $n = 8$. Bars show mean unobserved accuracy; the line shows mean slice condition number κ (right axis, log scale). Tick labels give observed pairs out of 64. (a) C_8: completing the generator column. (b) D_8: identity cross, one generator column, then added relator pairs.

J.4 Regularization sweep on S_4

Table 4 lists mean unobserved accuracy and the number of perfect runs (out of 5) for S_4 from `results/s4_results2.json`. The sweep shows the same transition as the passive benchmark: for every $\lambda \leq 1.5$, no run achieved full recovery at $m \leq 92$ (16.0%) under the tested training protocol, and all runs did at $m = 115$ (20.0%). Figure 4a compares these random masks with the presentation scaffold $S_{\min}$; panels (b–c) show training curves.

Table 4: S_4 sweep: mean unobserved accuracy (%) / perfect runs out of 5.

$\lambda \setminus m$ (fraction)	74 (13%)	86 (15%)	92 (16%)	97 (17%)	103 (18%)	115 (20%)
0.05	6.7 / 0	9.7 / 0	41.8 / 0	62.9 / 0	84.0 / 3	100.0 / 5
0.15	6.5 / 0	10.9 / 0	29.7 / 0	85.6 / 2	98.0 / 3	100.0 / 5
0.5	6.2 / 0	22.2 / 0	29.9 / 0	78.3 / 1	98.1 / 3	100.0 / 5
1.5	7.6 / 0	20.7 / 0	34.5 / 0	87.4 / 0	98.0 / 3	100.0 / 5
5.0	5.9 / 0	7.1 / 0	12.6 / 0	58.7 / 0	91.2 / 1	99.8 / 3

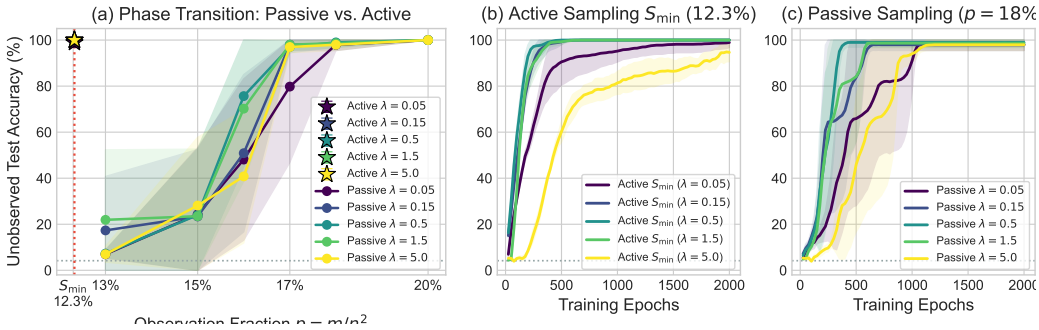


Figure 4: Active scaffold versus random masks on S_4. (a) Mean unobserved accuracy versus the observed fraction $p = m/n^2$ across regularization strengths λ; stars mark $S_{\min}$. (b–c) Training curves on $S_{\min}$ and on random masks with $p = 18\%$, respectively.

J.5 Active teaching on S_5

Historical 371-pair scaffold. On the non-Abelian permutation group S_5 ($n = 120$, table size 14,400), we trained on the scaffolds of Definition 5, with a transposition and a 5-cycle as generators and five relators (Appendix G), using Adam (learning rate 0.025, $\lambda = 0.5$, 600 epochs, 3 seeds). Table 5 (`results/s5_teaching_dimension_results.json`) shows that the identity cross achieves metric locking but stays at chance on the 14,161 unobserved pairs (chance 0.83%), confirming Proposition 8 on a 120-element group, and that a single generator column leaves τ unconstrained, while the 371-pair presentation scaffold and the 477-pair two-column scaffold both recover the table in all runs. The presentation scaffold uses 2.58% of the table, against 9.8% for passive recovery of S_5 (Table 2), demonstrating the quantitative gap between coordinated active teaching and uncoordinated random sampling.

Table 5: Active scaffolds on S_5 ($n = 120$), historical 371-pair runs, 3 seeds each. κ: slice condition number averaged over all slices; a few ill-conditioned slices can dominate this mean (see the 375-pair runs below).

Scaffold	Pairs	Fraction	Unobs. acc.	Recovered	κ	Epoch of $\geq 99\%$
S_{cross}	239	1.66%	0.85%	0/3	1.03	–
$S_{\text{cross}} \cup \sigma$-column	358	2.49%	3.45%	0/3	41.93	–
$S_{\min}$	371	2.58%	100.00%	3/3	14.91	275
S_{gen} (both columns)	477	3.31%	100.00%	3/3	1.24	100–125

Current 375-pair scaffold. The scaffold constructor in the supplied `dataset.py` yields a 375-pair $S_{\min}$ for S_5 (Appendix G), not the 371-pair scaffold of Table 5. Runs on the 375-pair scaffold are stored in `active_scaffold_benchmark_checkpoint.jsonl` (summarized in `active_scaffold_benchmark_results.json`): Gaussian initialization, Adam with learning rate 0.05, seeds 0–2, at $\lambda = 1.0$ and at $\lambda = 1.25$. All six runs reach 100% unobserved accuracy on the 14,025 unobserved pairs. The first evaluation epoch with unobserved accuracy $\geq 99\%$ is 350 in all three runs at $\lambda = 1.0$ and 400, 400, and 375 at $\lambda = 1.25$; the median slice condition numbers at the end of training are 1.015–1.017 at $\lambda = 1.0$ and 1.05–5.75 at $\lambda = 1.25$, whereas the means over slices are 2928, 5.7, and 15.2 at $\lambda = 1.0$ and 3.6, 19.6, and 47.1 at $\lambda = 1.25$. A few ill-conditioned slices therefore dominate the mean, which is the only statistic stored for the historical runs of Table 5. Perfect accuracy does not show that a trained solution has reached the exact common-scale geometry of Theorem 4. The two conditions are reported separately and are not pooled with the historical 371-pair runs.

K Optimization details

K.1 Dimension-consistent gradients and loss normalization

In standard implementations, the prediction loss is frequently evaluated as the batch mean $\mathcal{L}_{\text{mean}} = \frac{1}{|S|}\sum_{(a,b)\in S}\sum_c |\delta_{abc} - T_{abc}|^2$. Because the prediction tensor is normalized by the slice dimension, $T_{abc} = \frac{1}{n}\operatorname{Tr}(A_a B_b C_c)$, each gradient $\partial T_{abc}/\partial A_a$ scales as $1/n$. With $|S| = m = \Theta(n \log n)$, the effective gradient received by each slice scales as

$$\nabla_{A_a}\mathcal{L}_{\text{mean}} \sim \frac{1}{m}\sum_{b:(a,b)\in S}\frac{\partial T_{abc}}{\partial A_a} \sim \frac{1}{cn\ln n}\cdot(c\ln n)\cdot\frac{1}{n} = \mathcal{O}\left(\frac{1}{n^2}\right).$$

For large groups (e.g., $n = 120$), this factor dampens the gradient by more than 400-fold relative to $n = 6$, so fixed-step-size gradient descent slows down. Evaluating the unscaled sum $\sum_{(a,b)\in S}$ with gradient-norm clipping ($\|\mathbf{g}\| \le 1.0$) removes this damping.

K.2 Batch averaging versus set summation, and implementation settings

Deep learning frameworks often evaluate the sample mean $\frac{1}{|S|}\sum_{(a,b)\in S}$, whereas the theory uses the sum over S. Because $\operatorname{argmin}_\Theta \frac{1}{|S|}\mathcal{J}_{\lambda,\Omega_S} = \operatorname{argmin}_\Theta \mathcal{J}_{\lambda,\Omega_S}$, this change does not affect global or local minimizers or critical points; it scales the full-batch gradient by $|S|^{-1}$, which can be absorbed into the learning rate ($\eta_{\text{sum}} = \eta_{\text{mean}}/|S|$). The experiments use `loss_reduction="sum"` and `reg_norm="d"`, i.e. the regularizer $\sum_{(a,b)\in S} h_\Theta(a, b, a\circ b)$ of (3) with the normalized sensitivity $h_\Theta = n\, r_\Theta$, so the code parameter λ equals the λ of (5) and s_λ solves $s^3 + 2\lambda s = 1$ (`solve_s_lambda` in `models.py`). Slices are real-valued $n \times n$ matrices, and evaluation takes place every 25 epochs. The theory covers real slices: the scaled regular representation is real, so the infimum of (5) over real slices equals the complex infimum $|S|\phi_\lambda(s_\lambda)$ (Theorem 4), and every real global minimizer is a complex global minimizer, to which Theorem 4 and Corollary 11 apply. In the passive benchmark (`scripts/passive_sampling_exp.py`), initialization is Gaussian with standard deviation $n^{-1/2}$, and Adam uses learning rate 0.05, $\beta = (0.99, 0.999)$, gradient-norm clipping at 1, and at most 2000 epochs. The worker records the first evaluation epoch at which unobserved accuracy reaches 99% and stops 50 epochs later; a run that has not reached 99% stops at the first evaluation epoch ≥ 450 at which unobserved accuracy is within 5 percentage points of chance. Accuracy is reported at the stopping epoch, and both stopping rules read the unobserved accuracy. The S_5 scaffold runs of Table 5 (`scripts/run_s5_teaching_dimension_experiment.py`) use the model's default orthogonal initialization, learning rate 0.025, $\lambda = 0.5$, 600 epochs, and the trainer defaults for stopping (150 epochs after reaching 99%; near-chance check from epoch 500).

L Bernoulli comparison-graph connectivity

L.1 Notation

All logarithms are natural. In probability statements, $0 < \delta < 1$ and $0 \le p \le 1$ unless specified otherwise. An *observation mask* $\Omega \subseteq G^2$ is the set of observed input pairs (the set S of the main text); a cell means one such pair. A Bernoulli-p mask includes each cell of G^2 independently with probability p. We write

$$m = |\Omega|, \qquad \bar{m} = \mathbb{E}|\Omega| = pn^2. \tag{L.1}$$

The model is (1) with $A_a, B_b, C_c \in M_n(\mathbb{C})$, the space of complex $n \times n$ matrices; $X^\dagger$ denotes the conjugate transpose. A *matrix slice* is one matrix with a fixed label, such as A_a. "Untied" means that no equality between different families is imposed as an architectural condition; it is not a probabilistic independence assumption. $\mathrm{U}(n)$ denotes the unitary matrices. The local sensitivity h_Θ and its masked sum $\mathscr{H}_\Omega$ are (2) and (3) with $S = \Omega$; the minimum $3|\Omega|$ below is the same quantity as the minimum $3|S|$ of Theorem 3. The positive program (P_+), full-output constraints, anchors, and $(\mathrm{P}_{\mathrm{anc}})$ are as in Section 3.2.

In degree notation, put $d_A(g) = |\{b : (g,b) \in \Omega\}|$, $d_B(g) = |\{a : (a,g) \in \Omega\}|$, and $d_C(g) = |\{(a,b) \in \Omega : ab = g\}|$. Three-slot coverage (Definition A.1) is $d_A(g), d_B(g), d_C(g) > 0$ for every

$g \in G$. In the three-family model this is stronger than the shared-embedding condition $d_A(g) + d_B(g) + d_C(g) > 0$, because A_g, B_g, C_g are distinct unknowns. A *regular-gauge tuple* is a tuple of the form

$$A_a = X\rho_{\mathrm{reg}}(a)Y^\dagger, \qquad B_b = Y\rho_{\mathrm{reg}}(b)Z^\dagger, \qquad C_c = Z\rho_{\mathrm{reg}}(c)^\dagger X^\dagger \tag{L.2}$$

for fixed $X, Y, Z \in \mathrm{U}(n)$; by (6), it recovers the labeled tensor.

L.2 Deterministic synchronization without shared embedding

Theorem L.1 (Sampled equality rigidity)**.** *For every mask $\Omega \subseteq G^2$, the positive program* (4) *has minimum $3|\Omega|$, attained by the regular tuple*

$$A_a = \rho_{\mathrm{reg}}(a), \qquad B_b = \rho_{\mathrm{reg}}(b), \qquad C_c = \rho_{\mathrm{reg}}(c)^\dagger. \tag{L.3}$$

A feasible tuple is a global minimizer if and only if, for every $(a, b) \in \Omega$, the three slices A_a, B_b, C_{ab} involved in that observation are unitary and

$$A_a B_b C_{ab} = I_n. \tag{L.4}$$

Equivalently, on each observed pair,

$$A_a B_b = D_{ab}, \qquad D_c = C_c^\dagger, \tag{L.5}$$

with these three slices unitary. Under three-slot coverage, every slice in all three families is unitary.

Proof. For a feasible tuple and $(a, b) \in \Omega$, Lemma 1 gives $h_\Theta(a, b, ab) \geq 3\left|\frac{1}{n}\operatorname{Tr}(A_a B_b C_{ab})\right|^{4/3} = 3$. Summing yields $\mathscr{H}_\Omega \geq 3|\Omega|$. The tuple (L.3) is feasible by (6), and its three pairwise products at every observed triple are unitary. It attains the lower bound, proving existence of a global minimizer as well as the minimum value; no compactness or coercivity argument is needed. For any feasible tuple the excess above the minimum is a sum of nonnegative local differences, $\mathscr{H}_\Omega(\Theta) - 3|\Omega| = \sum_{(a,b)\in\Omega}\big(h_\Theta(a, b, ab) - 3\big)$. It is zero exactly when every summand is zero. The equality characterization in Lemma 1 yields unitarity and (L.4) on every observed triple. Multiplying by $C_{ab}^\dagger$ gives (L.5). The converse follows by reversing these steps. Three-slot coverage reaches each slice at least once, and hence makes all of them unitary. The empty-mask case is included: the objective is identically zero and the local characterization is vacuous. □

Corollary L.2 (Adding full-output constraints at observed pairs)**.** *Adding full-output constraints at any collection of observed pairs leaves the minimum equal to $3|\Omega|$. Every minimizer of the augmented program is also a minimizer of (P_+), so it satisfies every conclusion of Theorem L.1.*

Proof. The regular tuple obeys every such constraint and still attains $3|\Omega|$. The positive-only lower bound applies to the smaller feasible set. Thus both minima coincide, and any augmented-program minimizer is positive-feasible with the positive-program minimum value. □

From observed products to relative comparisons. The observed equations do not yet identify the three families with one common map. For fixed g, the label square in Figure 1a commutes because $(ga)b = g(ab)$, not because g and b commute. Only its horizontal pairs are assumed observed. Their matrix equations are $A_a B_b = D_{ab}$ and $A_{ga} B_b = D_{gab}$; their outputs belong to the D-family and cannot be replaced by A_{ab} or B_{ab}. Both equations nevertheless contain the same right factor B_b, which can be canceled as follows.

Lemma L.3 (Two observed equations give one relative comparison)**.** *Let $d \geq 1$ and suppose $A, B, D : G \to \mathrm{U}(d)$ satisfy $A_a B_b = D_{ab}$ for all $(a, b) \in \Omega$. If $g \in G$ and both $(a, b) \in \Omega$ and $(ga, b) \in \Omega$, then $A_{ga} A_a^\dagger = D_{g(ab)} D_{ab}^\dagger$.*

Proof. Associativity of the target group gives $(ga)b = g(ab)$. Therefore $D_{g(ab)} D_{ab}^\dagger = (A_{ga} B_b)(A_a B_b)^\dagger = A_{ga} B_b B_b^\dagger A_a^\dagger = A_{ga} A_a^\dagger$. Only the common right factor B_b was canceled. No two factor families were identified, and no matrices were commuted. □

The relative matrices in (L.7) implement left translation in their respective families:

$$P_g(a)A_a = A_{ga}, \qquad Q_g(x)D_x = D_{gx}.$$

These are the vertical operations in Figure 1b. A relative matrix may still depend on its starting label; the lemma gives only $P_g(a) = Q_g(ab)$. Under the shared-embedding restriction $A = D = R$, both definitions reduce to $R_{ga}R_a^\dagger$, but no such restriction is used here. To remove dependence on the starting label, we record the comparisons in a graph for each $g \neq e$; for $g = e$, all relative matrices are already I_d.

Definition L.4 (Relative-comparison graph)**.** Let $\Omega \subseteq G^2$ and $g \in G\backslash\{e\}$. The relative-comparison graph $\mathcal{K}_g(\Omega)$ is the bipartite graph with vertex classes G_A and G_D, each a copy of G, in which $a \in G_A$ and $x \in G_D$ are adjacent if and only if

$$(a, a^{-1}x) \in \Omega \quad \text{and} \quad (ga, a^{-1}x) \in \Omega. \tag{L.6}$$

For maps $A, D : G \to \mathrm{U}(d)$, the vertices $a \in G_A$ and $x \in G_D$ carry the matrices

$$P_g(a) = A_{ga}A_a^\dagger, \qquad Q_g(x) = D_{gx}D_x^\dagger. \tag{L.7}$$

Under the hypotheses of Lemma L.3, an edge $\{a, x\}$ certifies the observations (a, b) and (ga, b) with $b = a^{-1}x$, so adjacent vertices carry equal matrices.

Reading a comparison edge in either direction. The copies G_A, G_D are disjoint even when their group labels agree. An edge records two observed input pairs, not an additional observation at (a, x), and these are not the vertices or edges of the original input-pair graph. In the path of Figure 1c, put $b_1 = a_1^{-1}x_1$. The middle edge certifies $(a_1, b_1), (ga_1, b_1) \in \Omega$, hence

$$Q_g(x_1) = (A_{ga_1}B_{b_1})(A_{a_1}B_{b_1})^\dagger = A_{ga_1}A_{a_1}^\dagger = P_g(a_1).$$

Reading that edge from G_D to G_A uses symmetry of equality: it neither inverts the relative matrix nor replaces g by g^{-1}. Transitivity along the path gives $P_g(a_0) = Q_g(x_1) = P_g(a_1) = Q_g(x_2)$. Thus connectedness propagates equality of the relative matrices; it does not identify the raw slices or compose matrix labels along edges.

Definition L.5 (Good mask)**.** A mask $\Omega \subseteq G^2$ is good, written $\mathcal{E}_G(\Omega)$, if it has three-slot coverage (Definition A.1) and $\mathcal{K}_g(\Omega)$ is connected for every $g \neq e$. The property is monotone: if $\Omega \subseteq \Omega'$ and Ω is good, then Ω' is good. For a random mask, $\mathcal{E}_G(\Omega)$ denotes the corresponding event.

Lemma L.6 (Connected comparison graphs yield common left multipliers)**.** *Under the hypotheses of Lemma L.3, if $\mathcal{K}_g(\Omega)$ is connected, there exists $U_g \in \mathrm{U}(d)$ such that*

$$P_g(a) = Q_g(x) = U_g \qquad (a, x \in G). \tag{L.8}$$

Equivalently, the two families satisfy the left-translation relations

$$A_{ga} = U_g A_a, \qquad D_{gx} = U_g D_x \qquad (a, x \in G). \tag{L.9}$$

The same matrix U_g is a common left multiplier for both families: it implements left translation by g, independently of the starting labels a and x.

Proof. Along each graph edge the two assigned matrices agree by Lemma L.3. Any two vertices of a connected graph are joined by a finite path; transitivity of equality along the path makes their values equal. The common value is unitary, since every vertex value is unitary. Multiplication on the right by A_a or D_x converts (L.8) to (L.9). □

Lemma L.7 (Coverage supplied by comparison connectivity)**.** *If $\mathcal{K}_g(\Omega)$ is connected for at least one $g \neq e$, then every first-input label and every output label is covered. This conclusion need not hold for second-input labels.*

Proof. A connected graph on $2n \geq 4$ vertices has no isolated vertices. For every left vertex a, some incident edge certifies an observed pair whose first coordinate is a. For every right vertex x, some incident edge certifies $(a, a^{-1}x) \in \Omega$, whose product is x. There is no corresponding vertex indexed by the second-input label b, so this argument gives no second-slot coverage. □

Comparison connectivity covers the first-input and output labels, but supplies no vertex indexed by the second-input label b. The separate column-coverage hypothesis below provides an observed equation for each B_b. Here synchronization means expressing the three families through one representation and fixed unitary coordinate changes; it is a conclusion about a parameter tuple, not about the evolution of a training algorithm.

Theorem L.8 (Three-family synchronization and full-law recovery)**.** *Let G be a finite group of order $n \geq 2$ and $d \geq 1$. Suppose $A, B, D : G \to \mathrm{U}(d)$ satisfy $A_a B_b = D_{ab}$ for all $(a, b) \in \Omega$. Assume that every input column is nonempty and that $\mathcal{K}_g(\Omega)$ is connected for every $g \neq e$. Then there are a representation $U : G \to \mathrm{U}(d)$ and fixed $P, Q \in \mathrm{U}(d)$ such that*

$$A_a = U_a P, \qquad B_b = P^\dagger U_b Q, \qquad D_c = U_c Q \quad (a, b, c \in G). \tag{L.10}$$

In particular $A_a B_b = D_{ab}$ on all of G^2.

Proof. For each $g \neq e$, Lemma L.6 gives a unitary U_g satisfying (L.9). Put $U_e = I_d$; the same left-translation relations then hold for $g = e$. Define $P = A_e$ and $Q = D_e$. Setting $a = e$ and $x = e$ in the left-translation relations yields $A_g = U_g P$ and $D_g = U_g Q$ for every g. For arbitrary $g, h \in G$, applying $A_{ga} = U_g A_a$ at $a = h$ gives $U_{gh} = A_{gh} A_e^\dagger = U_g A_h A_e^\dagger = U_g U_h$. This also covers products equal to e, because $A_e A_e^\dagger = I_d = U_e$. Thus U is a representation. Fix $b \in G$. By column coverage choose a with $(a, b) \in \Omega$. The observed equation and the formulas for A and D give $B_b = A_a^\dagger D_{ab} = P^\dagger U_a^\dagger U_{ab} Q = P^\dagger U_b Q$. This formula no longer depends on the chosen observed row. Finally, for any pair, observed or not, $A_a B_b = U_a P P^\dagger U_b Q = U_{ab} Q = D_{ab}$. □

The normal form also explains why a shared representation can be obtained without imposing shared embedding. Since $C_c = Q^\dagger U_c^\dagger$, the unitary gauge transformation

$$(A_a, B_b, C_c) \longmapsto (A_a P^\dagger,\ P B_b Q^\dagger,\ Q C_c) = (U_a, U_b, U_c^\dagger)$$

synchronizes the families. Shared embedding is therefore a conclusion after an allowed gauge transformation.

By Lemma L.7, the good-mask condition of Definition L.5 is equivalent to column coverage together with the connectivity conditions; we retain the symmetric three-slot formulation when estimating its probability. For a fixed mask it is a deterministic condition; for a random mask it defines an event.

Corollary L.9 (All positive-only minimizers on a good mask)**.** *If $\mathcal{E}_G(\Omega)$ holds, every global minimizer of (P_+) has the form* (L.10) *with $d = n$. Conversely every tuple of that form, with $C_c = Q^\dagger U_c^\dagger$, is a global minimizer of (P_+).*

Proof. The forward direction combines Theorems L.1 and L.8. Conversely the representation law makes the positive triple product I_n at every pair, and all slices are unitary. Every local sensitivity is 3, so the tuple attains $3|\Omega|$. The event is independent of the matrices, and hence the characterization holds simultaneously for all minimizers. □

Lemma L.10 (One output vector determines the regular character)**.** *Let $U : G \to \mathrm{U}(n)$ be a representation. Suppose that for one $g_0 \in G$,*

$$\frac{1}{n} \operatorname{Tr}(U_{g_0} U_c^\dagger) = \mathbf{1}_{\{g_0 = c\}} \quad (c \in G). \tag{L.11}$$

Then

$$\operatorname{Tr}(U_h) = n \mathbf{1}_{\{h = e\}} \quad (h \in G), \tag{L.12}$$

and U is unitarily equivalent to ρ_{reg}.

Proof. This is Theorem 7 (i)–(ii) applied to $R = U$ with $a_0 b_0 = g_0$; see Appendix B.3. □

Theorem L.11 (Complete labeled recovery on a good mask)**.** *Assume $\mathcal{E}_G(\Omega)$, and augment (P_+) by imposing full-output constraints at any nonempty collection of observed pairs (possibly a single pair). Then the resulting program has minimum $3|\Omega|$, and every global minimizer has the regular-gauge form* (L.2) *and satisfies*

$$T_{abc}(\Theta) = \mathbf{1}_{\{ab = c\}} \quad (a, b, c \in G). \tag{L.13}$$

Conversely, every regular-gauge tuple is a global minimizer of every such augmented program.

Proof. The minimum and local rigidity follow from Corollary L.2. By Theorem L.8, write a minimizer in the form (L.10), so $A_aB_bC_c = U_aPP^\dagger U_bQQ^\dagger U_c^\dagger = U_aU_bU_c^\dagger$. For an anchored pair with product g_0, its full-output constraints are exactly (L.11). By Lemma L.10, choose V so that $U_g = V\rho_{\mathrm{reg}}(g)V^\dagger$. Setting $X = V$, $Y = P^\dagger V$, $Z = Q^\dagger V$ gives (L.2), with every factor in the correct order. Formula (6) gives (L.13). Conversely every regular-gauge tuple has the complete target tensor and local sensitivities 3. It satisfies all imposed full-output constraints and attains the minimum. □

L.3 Bernoulli comparison-graph connectivity

For $g \neq e$, a comparison edge uses two distinct cells, giving marginal probability p^2. Different edges may reuse cells, so the diagram does not make their indicators independent. The proof works with the original Bernoulli mask: for each fixed cut, bounded overlap supplies many candidate edges whose required cells are disjoint. Their absence can then be bounded using independence before taking the union bound over cuts and translations.

Write $\xi_{u,v} := \mathbf{1}_{\{(u,v)\in\Omega\}}$ for the independent Bernoulli-p inclusion indicators of the mask. For a fixed $g \neq e$, the indicator of the candidate edge joining $a \in G_A$ to $x \in G_D$ is

$$Z_{a,x} = \xi_{a,a^{-1}x}\,\xi_{ga,a^{-1}x}. \tag{L.14}$$

A candidate edge is a possible edge indexed by (a, x), whether or not its two required cells are observed.

Lemma L.12 (Bounded overlap of comparison edges)**.** *Let Ω be a Bernoulli-p mask and fix $g \neq e$. Then each candidate edge is present with probability p^2, and each observation cell is used by at most two candidate edges. Moreover, for every fixed nonempty set F of candidate edges,*

$$\mathbb{P}(\textit{no edge of } F \textit{ is present}) \le (1-p^2)^{|F|/3} \le e^{-p^2|F|/3}. \tag{L.15}$$

Here F is chosen independently of the realized mask.

Proof. The two cells in (L.14) are distinct because $ga \neq a$. Their inclusion indicators are independent, so the edge has probability p^2. A fixed cell (u, v) can occur as the first required cell only in the candidate edge (u, uv), and as the second only in $(g^{-1}u, (g^{-1}u)v)$. Thus each cell is used at most twice, and each candidate edge shares a required cell with at most two other candidates. Starting from F, repeatedly select one remaining candidate edge and discard it and all remaining candidates sharing either of its required cells. Each step discards at most three candidates. When none remain, the selected set F_0 has $|F_0| \ge |F|/3$, and different edges in F_0 use disjoint sets of observation cells. The selection uses only F and the fixed cell locations, not the realized mask. Hence their occurrence indicators are jointly independent Bernoulli-p^2 variables. Absence of every edge of F implies absence of every selected edge, so $\mathbb{P}(\text{no edge of } F \text{ is present}) \le (1-p^2)^{|F_0|} \le (1-p^2)^{|F|/3} \le e^{-p^2|F|/3}$. When $g^2 = e$, the distinct candidates (a, x) and (ga, gx) use the same two cells in reverse order. The selection never retains both, so they are not counted as independent. □

Lemma L.13 (Direct comparison-graph connectivity bound)**.** *Let Ω be a Bernoulli-p mask and fix $g \neq e$. Then*

$$\mathbb{P}(\mathcal{K}_g(\Omega) \textit{ disconnected}) \le \sum_{k=1}^{n} \binom{2n}{k} e^{-np^2k/6}. \tag{L.16}$$

If $2ne^{-np^2/6} \le 1/2$, then

$$\mathbb{P}(\mathcal{K}_g(\Omega) \textit{ disconnected}) \le 4ne^{-np^2/6}. \tag{L.17}$$

Proof. If a graph on $2n$ vertices is disconnected, one of its components has size k with $1 \le k \le n$. Consider a fixed vertex set W of size k, containing i left vertices and j right vertices, with $i + j = k$. A crossing edge joins W to its complement. Let F_W be the set of candidate crossing edges. Its size is $|F_W| = i(n-j) + j(n-i) = nk - 2ij \ge nk - \frac{k^2}{2} \ge \frac{nk}{2}$, using $ij \le k^2/4$ and $k \le n$. By Lemma L.12, $\mathbb{P}(\text{no edge crosses } W) \le e^{-p^2|F_W|/3} \le e^{-np^2k/6}$. The disjoint-cell subset is selected separately for each fixed cut; it need not itself form a connected spanning graph. We need not require

internal connectivity of W, because ignoring it only enlarges the event. There are $\binom{2n}{k}$ candidate vertex sets of size k. A union bound proves (L.16). Writing $t = 2ne^{-np^2/6}$ and using $\binom{2n}{k} \le (2n)^k$, the right side is at most $\sum_{k=1}^{n} t^k$. For $t \le 1/2$ this is at most $t/(1-t) \le 2t = 4ne^{-np^2/6}$. □

Lemma L.14 (Three-slot coverage probability)**.** *Let Ω be a Bernoulli-p mask. Then*

$$\mathbb{P}(\textit{three-slot coverage fails}) \le 3n(1-p)^n \le 3ne^{-pn}. \tag{L.18}$$

Proof. A fixed row and a fixed column each contain n cells. For a fixed output g, its cells are $(a, a^{-1}g)$ for $a \in G$, also n distinct cells. Each one of these $3n$ slot-label classes is empty with probability $(1-p)^n$. Union-bound their failure events; independence between different slot-label classes is neither true in general nor needed. □

Proposition L.15 (Probability of the good-mask condition)**.** *Let Ω be a Bernoulli-p mask. If $2ne^{-np^2/6} \le 1/2$, then*

$$\mathbb{P}\big(\mathcal{E}_G(\Omega)^c\big) \le 4n(n-1)e^{-np^2/6} + 3ne^{-pn}. \tag{L.19}$$

In particular, for $0 < \delta < 1$, the explicit condition

$$\boxed{p^2 n \ge 6 \log \frac{8n^2}{\delta}} \tag{L.20}$$

implies $\mathbb{P}(\mathcal{E}_G(\Omega)) \ge 1 - \delta$.

Proof. Apply Lemma L.13 and union-bound over the $n-1$ nonidentity elements. Add Lemma L.14 to obtain (L.19). Neither independence between the different comparison graphs nor independence between connectivity and coverage is required. To check the explicit constants, let $L = \log(8n^2/\delta)$. Condition (7) yields $np^2/6 \ge L$. In particular $2ne^{-np^2/6} \le \frac{\delta}{4n} \le \frac{1}{2}$, so the preceding connectivity estimate applies. Its graph-failure contribution is at most $4n(n-1)e^{-L} = \frac{\delta(n-1)}{2n} < \frac{\delta}{2}$. Also $pn \ge p^2 n \ge 6L \ge L$ for $p \le 1$, so the coverage contribution is at most $3ne^{-pn} \le 3ne^{-L} = \frac{3\delta}{8n} \le \frac{3\delta}{16}$. Their sum is less than δ. □

L.4 Proofs of the main passive theorem and its corollaries

Theorem 9 (Exact partial recovery without shared embedding)**.** *Let G be a fixed finite group of order $n \ge 2$, let $0 < \delta < 1$, and let $\Omega \subseteq G^2$ be a Bernoulli-p mask. Use the three-family model* (1)*, the masked regularizer* (3)*, and $n \times n$ complex slices. Then, for every mask, the positive constrained program* (P_+) *has minimum $3|\Omega|$. If* (7) *holds, then there is an event of probability at least $1 - \delta$ on which the following statements hold simultaneously:*

1. *Every positive-only global minimizer has unitary slices and is of the form $A_a = U_a P$, $B_b = P^\dagger U_b Q$, $C_c = Q^\dagger U_c^\dagger$, for a representation $U : G \to \mathrm{U}(n)$ and fixed unitaries P, Q. In particular, $A_a B_b = C_{ab}^\dagger$ at every input pair.*
2. *The mask is nonempty. At any observed pair selected after seeing the mask, adding its full-output constraints leaves the minimum at $3|\Omega|$ and forces every augmented-program global minimizer to be a regular-gauge tuple.*
3. *The anchored minimizers recover the entire labeled tensor: $T_{a,b,c}(\Theta) = \mathbf{1}_{\{ab=c\}}$ for all triples. The same conclusion holds when full-output constraints are imposed at any nonempty set of observed pairs, in particular at all of them.*

Proof of Theorem 9. The minimum for an arbitrary mask is given by Theorem L.1. By Proposition L.15, the event that $\mathcal{E}_G(\Omega)$ holds has probability at least $1 - \delta$. On that event Corollary L.9 gives the first assertion, and coverage implies nonemptiness. For every possible choice of an observed anchor, Theorem L.11 gives the remaining conclusions. These are deterministic consequences of the same good-mask event, so no union bound over matrices, minimizers, or choices of anchor is required. □

Proof of Corollary 10. Since $p = \bar{m}/n^2$, one has $p^2 n = \bar{m}^2/n^3$, so (7) reads $\bar{m}^2 \geq 6n^3 \log(8n^2/\delta)$. Both sides of the resulting square-root inequality are nonnegative. □

Proof of Corollary 11. On $\mathcal{E}_G(\Omega)$ the mask has three-slot coverage, so Theorem 4 applies to every exact global minimizer $\widehat{\Theta}$ of $\mathcal{J}_{\lambda,\Omega}$: the normalized tuple $\widetilde{\Theta} = s_\lambda^{-1}\widehat{\Theta}$ has unitary slices, satisfies $\widetilde{A}_a \widetilde{B}_b \widetilde{C}_{ab} = I_n$ at every observed pair (proof of Theorem 4, step (iii)), and satisfies $T_{\widetilde{\Theta}}(a,b,c) = s_\lambda^{-3} T_{\widehat{\Theta}}(a,b,c) = \mathbf{1}_{\{ab=c\}}$ for every observed pair (a,b) and every c (steps (ii) and (iv)). Hence $\widetilde{\Theta}$ is feasible for the program with full-output constraints at all observed pairs, and its value $\mathscr{H}_\Omega(\widetilde{\Theta}) = 3|\Omega|$ is that program's minimum (Corollary L.2); so $\widetilde{\Theta}$ is a global minimizer of it. Theorem L.11 gives $T_{\widetilde{\Theta}} = \mathbf{1}_{\{ab=c\}}$ on G^3, and therefore $T_{\widehat{\Theta}} = s_\lambda^3 \mathbf{1}_{\{ab=c\}}$. Since $s_\lambda^3 > 0$, the argmax over c is ab for every pair. The probability statement follows from Proposition L.15; the conclusion is deterministic on the event, so it holds for all minimizers simultaneously. □

Corollary L.16 (Fixed-size masks)**.** *Let Ω_m be a uniformly random set of m distinct pairs of G^2, as in the experiments of Section 5, and put $p = m/n^2$. Then $\mathbb{P}(\mathcal{E}_G(\Omega_m)^c) \leq 2\,\mathbb{P}(\mathcal{E}_G(\Omega)^c)$ for a Bernoulli-p mask Ω. In particular, if $p^2 n \geq 6\log(8n^2/\delta)$, the deterministic conclusions of Theorem 9 and Corollary 11 hold for Ω_m with probability at least $1 - 2\delta$.*

Proof. The complement D of the good-mask event is decreasing: if $\Omega \subseteq \Omega'$ and $\Omega' \in D$, then $\Omega \in D$ (Definition L.5). Removing a uniformly random element from a uniform $(k+1)$-subset of G^2 gives a uniform k-subset contained in it, so $q_k := \mathbb{P}(\Omega_k \in D)$ is nonincreasing in k. Conditional on $|\Omega| = k$, a Bernoulli-p mask is a uniform k-subset. Hence $\mathbb{P}(\Omega \in D) = \sum_k \mathbb{P}(|\Omega| = k)\, q_k \geq \mathbb{P}(|\Omega| \leq m)\, q_m \geq q_m/2$, because $|\Omega|$ is binomial with integer mean $n^2 p = m$, which is then also its median (Kaas & Buhrman, 1980). The last claim follows from Proposition L.15, since the conclusions of Theorem 9 and Corollary 11 are deterministic on the good-mask event. □

L.5 Associativity squares: from local filling to connected comparisons

The comparison-graph argument can be motivated by first considering how observed multiplication laws propagate around an associativity square. For unitary slices in the shared-embedding model, three observed sides force the fourth law. With only the two horizontal sides observed, one can no longer conclude either vertical law, but cancellation still equates the corresponding relative matrices. This weaker conclusion is the starting point of the relative-comparison argument: it remains valid for untied factor families, and connected paths propagate the equalities into common left multipliers.

The shared-embedding model assigns a single matrix R_g to each $g \in G$ across all three slots, with prediction tensor $T_R(a,b,c) = \frac{1}{n}\operatorname{Tr}(R_a R_b R_c^\dagger)$. We first describe how three observed laws fill a missing law, then connect the two-observation comparison to Figure 1c.

Definition L.17 (Right associativity witness)**.** Let $\Omega \subseteq G^2$ be an observation mask. For a target pair $(a,b) \in G^2$, an element $c \in G$ is a **right associativity witness** if the three pairs (b,c), (a,bc), (ab,c) all belong to Ω. We write $\mathcal{W}_\Omega(a,b) := \{c \in G : (b,c), (a,bc), (ab,c) \in \Omega\}$.

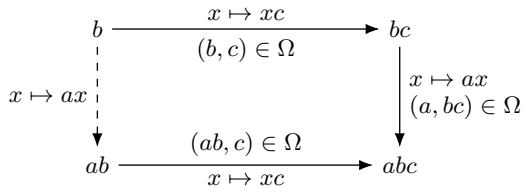


Figure 5: A one-slit associativity square in the shared-embedding model. The two horizontal pairs and the right vertical pair are observed; the dashed left side is the law at (a,b) to be deduced. The square commutes because $a(bc) = (ab)c$. The three observed matrix laws, together with invertibility of R_c, force $R_a R_b = R_{ab}$.

In Figure 5, the horizontal arrows act by right multiplication by c, and the vertical arrows by left multiplication by a. The group-label square commutes by associativity, independently of any learned

matrix law. In the shared-embedding model, the three observed pairs give $R_bR_c = R_{bc}$, $R_aR_{bc} = R_{abc}$, and $R_{ab}R_c = R_{abc}$. The defect identity below shows how they force the law on the dashed side.

Filling a law versus comparing translations. Retaining only the two horizontal laws in Figure 1a gives $R_{ga}R_a^\dagger = R_{gab}R_{ab}^\dagger$, without identifying either relative matrix with R_g. In the untied model, the equality is $P_g(a) = Q_g(ab)$, propagated along connected comparison graphs as in Figure 1c. One-step witnesses use three observed cells (allowing coincidences); comparison edges for $g \neq e$ use two distinct cells. The respective sufficient conditions involve np^3 and np^2, giving the orders $n^{5/3}(\log n)^{1/3}$ and $n^{3/2}(\log n)^{1/2}$. The comparison-graph argument gives a smaller asymptotic sufficient bound and applies to the three-family model used in the experiments; the witness argument applies to the shared-embedding model.

Lemma L.18 (Associativity-square defect identity)**.** *Let $d \geq 1$ and let $R : G \to M_d(\mathbb{C})$ be any matrix-valued map with multiplicative defect $\Delta(x, y) := R_xR_y - R_{xy}$. Then for every $a, b, c \in G$, $\Delta(a, b)R_c = R_a\Delta(b, c) + \Delta(a, bc) - \Delta(ab, c)$.*

Proof. Expanding the right-hand side, $R_a(R_bR_c - R_{bc}) + (R_aR_{bc} - R_{a(bc)}) - (R_{ab}R_c - R_{(ab)c}) = R_aR_bR_c - R_{ab}R_c = (R_aR_b - R_{ab})R_c$, where $R_{a(bc)} = R_{(ab)c}$ by associativity $a(bc) = (ab)c$ (`associativity_defect_identity`). □

This mechanism generalizes the commutative parallelogram closures observed by Liu et al. (2022) to non-Abelian groups.

Exact defect propagation. At any exact global minimizer attaining the Matrix AM–GM floor, every observed multiplication triple satisfies local locking: $\Delta(x, y) = 0$ for all $(x, y) \in \Omega$. If a missing pair $(a, b) \notin \Omega$ admits a witness $c \in \mathcal{W}_\Omega(a, b)$, then the three defects on the right-hand side vanish, so $\Delta(a, b)R_c = 0$. Because c appears in observed triples, R_c is unitary, hence invertible, which forces $\Delta(a, b) = 0$, i.e. $R_aR_b = R_{ab}$. Missing entries are filled algebraically without word induction along paths.

In the shared-embedding model, write $h_R(a, b) := \frac{1}{n}\big(\|R_aR_b\|_F^2 + \big\|R_bR_{ab}^\dagger\big\|_F^2 + \big\|R_{ab}^\dagger R_a\big\|_F^2\big)$. The *shared-embedding positive program* minimizes $\sum_{(a,b)\in\Omega} h_R(a, b)$ subject to $T_R(a, b, ab) = 1$ for $(a, b) \in \Omega$. By Lemma 1 with $(X, Y, Z) = (R_a, R_b, R_{ab}^\dagger)$, each term is at least 3, with equality if and only if R_a, R_b, R_{ab} are unitary and $R_aR_b = R_{ab}$; the regular representation attains $3|\Omega|$. Hence every global minimizer satisfies $\Delta = 0$ on Ω, with unitary R_g for every label g that occurs in an observed pair.

Theorem L.19 (Sub-quadratic sample complexity for one-step associativity completeness)**.** *Let G be a finite group of order $n \geq 2$, let $\Omega \subseteq G^2$ be a Bernoulli-p random mask, and let $\delta \in (0, 1)$. If $p^3n \geq 3\log(n^2/\delta)$, equivalently if the expected number of observed input pairs satisfies*

$$\bar{m} = \mathbb{E}|\Omega| = pn^2 \geq 3^{1/3}n^{5/3}\Big(\log\frac{n^2}{\delta}\Big)^{1/3} = \mathcal{O}\big(n^{5/3}(\log n)^{1/3}\big),$$

then there is an event of probability at least $1 - \delta$ on which:

1. ***One-step witness completeness:*** *every pair $(a, b) \in G^2$, in particular every missing pair, has at least one right associativity witness in Ω;*
2. ***Representation-law recovery:*** *every global minimizer of the shared-embedding positive program has unitary slices and satisfies $R_aR_b = R_{ab}$ for all $a, b \in G$, forming a unitary representation of G;*
3. ***Anchored Cayley recovery:*** *if full-output constraints are added at one observed pair, the minimum remains $3|\Omega|$, and every global minimizer is unitarily equivalent to the regular representation and recovers the full Cayley tensor on all n^3 triples.*

Proof. (1) Fix $(a, b) \in G^2$. For $c \in G$, the witness event W_c requires the cells $x_c = (b, c)$, $y_c = (a, bc)$, and $z_c = (ab, c)$. At most three distinct cells are involved ($x_c = z_c$ when $a = e$;

$y_c = z_c$ when $b = e$; all three coincide when $a = b = e$), so $\mathbb{P}(W_c) \geq p^3$. Each of the maps $c \mapsto x_c, y_c, z_c$ is injective. For $c \neq c'$, a cell of W_c equals a cell of $W_{c'}$ only if $x_c = y_{c'}$ or $y_c = x_{c'}$: the remaining coincidences $x_c = z_{c'}$, $z_c = x_{c'}$, $y_c = z_{c'}$, $z_c = y_{c'}$ require $a = e$ or $b = e$, and then force $c = c'$. The two possible coincidences require $a = b$ and $c' \in \{b^{-1}c, bc\}$. Thus each W_c shares a cell with at most two other events. Selecting greedily as in Lemma L.12 yields a set $C \subseteq G$ with $|C| \geq n/3$ whose events use pairwise disjoint cells and are therefore independent, so $\mathbb{P}(\text{no witness for } (a,b)) \leq \prod_{c \in C}(1-p^3) \leq e^{-p^3 n/3}$. A union bound over the n^2 pairs gives failure probability at most $n^2 e^{-p^3 n/3} \leq \delta$.

(2) On this event, every $g \in G$ occurs as a first input of an observed pair (take $b = g$ and a witness c, so $(g, c) \in \Omega$); hence every global minimizer has unitary R_g for all g and $\Delta = 0$ on Ω. For a missing pair (a, b) with witness c, Lemma L.18 gives $\Delta(a, b)R_c = 0$, and R_c is invertible, so $\Delta(a, b) = 0$. Thus $R_a R_b = R_{ab}$ on G^2, and $R_e = I$ because $R_e^2 = R_e$ with R_e invertible.

(3) The regular representation satisfies every full-output constraint and attains $3|\Omega|$, so the minimum is unchanged and every augmented minimizer is a minimizer of the positive program. By (2) it is a unitary representation, and the full-output constraints at the anchored pair are the hypothesis of Theorem 7, which gives unitary equivalence to ρ_{reg} and $T_R = \delta$ on G^3. The event in (1) depends only on Ω, so (2) and (3) hold simultaneously for all minimizers. □

The deterministic propagation step—that right associativity witnesses on unobserved pairs force the representation law across the entire group—is mechanically verified in Lean 4 (Appendix M); the probabilistic witness-coverage bound is proved analytically.

M Formal verification

Formal Verification and Scope. For finite groups in the complex-matrix model, the accompanying Lean 4 development formalizes metric locking, character recovery, canonical-column identifiability, and presentation-scaffold recovery under explicit presentation and inverse-consistency assumptions. It also formalizes passive Bernoulli recovery, including three-slot coverage and comparison-graph connectivity, finite-positive- regularization recovery, and structural counterexamples. Fixed-size and with-replacement probability bounds, the full isotope-transport extension, and the sharper shared-embedding witness bound stated in the paper remain outside this formalized scope; a weaker shared-embedding bound is formalized. The Lean sources, detailed correspondence between manuscript statements and formal declarations, axiom-dependency audit, and recorded verification logs are available in the public repository (`https://anonymous.4open.science/r/HyperCube_Group_LEAN-8E63/`).

N Extended related work

Grokking and sample complexity. Learning symbolic relations $a \circ b = c$ over a finite vocabulary of size n is a fundamental canonical setting for studying algebraic generalization and exact operator discovery in deep learning. Recent empirical studies have shown that standard architectures (such as transformers and multilayer perceptrons (MLPs)) exhibit *grokking* on modular arithmetic: training loss drops to zero rapidly via memorization, followed hundreds of epochs later by a delayed generalization phase transition (Power et al., 2022; Nanda et al., 2023; Gromov, 2023). While recent theoretical analysis by Mohamadi et al. (2024) established an $\widetilde{\mathcal{O}}(n^{5/3})$ sample complexity bound (where $\widetilde{\mathcal{O}}$ suppresses polylogarithmic factors) for (the generalization phase of) two-layer networks with quadratic activations (QMLP) on Abelian modular arithmetic ($\mathbb{Z}_n$), sample complexity bounds for general finite groups—especially non-Abelian groups—have remained entirely open.

Low-Rank Matrix Completion. Classical guarantees for matrix completion rely on two core components: a low-rank prior to restrict degrees of freedom, and incoherence conditions to ensure the matrix's information is not concentrated in a few unobservable entries (Candès & Recht, 2009; Ge et al., 2016). Our result for group table completion mirrors this architecture by replacing the low-rank prior with an associativity prior, and incoherence conditions with the Latin-square property of quasigroups (van Lint & Wilson, 2001, Ch. 17). Under uniform input sampling, this Latin property ensures uniform row, column, and output marginals, yielding three-slot coverage at the

coupon-collector rate (Theorem A.7)—bypassing the need for geometric incoherence assumptions. Exact recovery is attained by enforcing the associativity prior, which our model achieves either by propagating equalities derived from observed product pairs under passive sampling (Theorem 9), or by tracing the defining relators of an active presentation scaffold (Theorem 6). In either case, a complete output fiber at a single observed pair provides the necessary anchor to select the regular representation and recover the full table (Theorem 7).