A SPECTRAL THEORY OF COMPOSITIONAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

How does compositional reasoning emerge during learning? We address this question by mathematically analyzing the learning dynamics of deep linear networks. We train these networks in structured synthetic environments and derive a theory linking the structure of experience to compositional learning. Our theory predicts when compositional inferences emerge, whether they are identifiable from the available evidence, and how new linking evidence can rapidly unlock previously unavailable inferences. These results provide a qualitative explanation for several phenomena observed in human cognition. They account for why a composition can fail despite knowing its premises, why similar compositions can emerge at different times, and how a single linking fact can suddenly enable many new inferences. Taken together, these findings establish a mathematical link between the statistical structure of experience and the development of compositional reasoning.

1 INTRODUCTION

Compositional reasoning is fundamental to human intelligence. It allows us to combine known relations into new ones, extending our knowledge beyond direct experience (Frankland & Greene, 2020; Halford et al., 2010). For example, if we know that $a > b$ and $b > c$, we can infer that $a > c$ without observing that fact. This capacity supports reasoning across domains, from interpreting social and spatial relationships to understanding language and solving unfamiliar problems (Behrens et al., 2018).

Compositional reasoning, however, does not arise uniformly across contexts. In associative-inference tasks, individuals can learn overlapping relational premises such as A–B and B–C yet still remain poor at grasping the unseen A–C composition (Zeithamova & Preston, 2010; Kumaran & McClelland, 2012). How those relations are experienced also matters. When learning cleanly isolates the structural components of premises, compositional generalization improves and can emerge through intermediate partial solutions (Dekker et al., 2022). By contrast, when two relational structures have already been learned separately, only a small amount of linking evidence can rapidly unlock many compositional inferences (Nelli et al., 2023). Thus, similar premise knowledge can result in persistent compositional failure, gradual generalization, or sudden inference depending on how relational evidence is organized and presented. So, what determines which of these trajectories occurs?

Mathematical theories of learning offer a way to explain these differences. In deep linear networks, the training data can be decomposed into separate components, and each component is learned at a speed set by how strongly the data express it (Saxe et al., 2014; 2019; Lampinen & Ganguli, 2019). These components and their strengths make up the spectrum of the training data, which determines which parts of a representation are learned and when. These analyses concern outputs that a network is trained to produce, whereas a compositional inference is, by definition, never trained. However, a composition combines relations that are each learned from training facts, so it can only become available once those relations are learned. Its timing should therefore depend on how strongly the training facts constrain those relations, just as the strength of a component sets how quickly it is learned. If the training facts leave one of these relations undetermined, further training on the same facts should not be able to settle it, and the composition remains unavailable until a new fact is added. When several compositions depend on the same undetermined relation, a single new fact that determines it can make all of them available at once. If this rule holds, gradual

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.

generalization, persistent compositional failure, and sudden inference could all arise from the same learning dynamics.

We test this account in synthetic relational worlds with known additive laws, using shallow and deep linear neural networks. Our contributions are as follows:

- In Section 4.2, we give a criterion for whether the observed facts uniquely determine a compositional law, even when the full embedding remains undetermined (Proposition 4.1).
- In Sections 4.3–4.4, we derive closed-form shallow learning trajectories as sums of decaying spectral modes (Proposition 4.2) and numerically integrate the coupled dynamics of deeper networks to predict compositional retrieval.
- In Section 5, we test these predictions in three experiments and three network depths. Across independently generated worlds, the predictions track the trained networks' retrieval and geometry, showing how evidence governs law-specific emergence and how linking facts can resolve previously unavailable inferences.

2 Related Work

Relational representations. Relational embeddings can encode compositional rules by making a sequence of relations act like a single relation. Translation models give this an additive form by representing relations as vector displacements (Bordes et al., 2013). More broadly, training can produce algebraic relationships between relation representations that reflect compositional patterns in the data (Guu et al., 2015; Sun et al., 2019). These learned relationships are the object of our analysis. In additive worlds, we characterize which observed facts require composed displacements to agree with a single relation and predict how that agreement develops during learning.

Structural identifiability. Training examples can admit several interpretations that disagree on unseen combinations. Identification-based theories establish conditions under which training coverage and architectural constraints make compositional structure recoverable without observing every combination (Wiedemer et al., 2023; Schug et al., 2024). Thus, they show that, for relational inference, different embeddings can satisfy the same observations and enforce the same compositional law while disagreeing in others. We build on this knowledge in Proposition 4.1, where we characterize when a particular law should hold across all embeddings consistent with the evidence.

Compositional learning dynamics. The structure of training evidence influences both the order and course of generalization. Spectral analyses connect learning times to the recovery of structured representations (Saxe et al., 2019; Liu et al., 2022), while interactions between learning directions can produce temporary reversals in compositional performance (Yang et al., 2025). Our analysis contributes to this literature by deriving the mean learning dynamics from the relational training objective and using it to follow individual laws through learning. Reconstructing the evolving dynamics yields predictions for their retrieval trajectories, including after linking evidence changes the constraints.

Hidden progress and sudden emergence. Successful generalization can become visible only after substantial internal progress. Mechanistic analyses of grokking and studies of concept learning reveal developing computational structure and capabilities that ordinary performance measures initially conceal (Nanda et al., 2023; Park et al., 2024). Their findings motivate distinguishing progress toward an inference from successful retrieval. Our geometric error measures how closely a query approaches its target, including while retrieval remains incorrect. Predicting this error alongside accuracy lets us examine how changes in the learned geometry precede and produce observable compositional success.

Knowledge integration. Learning a new fact need not make its consequences available. Knowledge-editing benchmarks demonstrate that language models can reproduce an edited fact while failing to update answers that depend on it (Zhong et al., 2023; Cohen et al., 2024). Human knowledge-assembly experiments show how consequential such propagation can be, with limited linking information enabling many inferences across separately learned structures (Nelli et al.,

2023). Our linking experiment isolates this problem in a consistent additive system. We identify the shared ambiguity removed by a linking fact and predict how its consequences become retrievable, including temporary disruption of previously learned premises.

3 Experimental Framework

3.1 Relational worlds and learning model

Our worlds consist of entities connected by directed relations (Fig. 1). A fact $a \xrightarrow{r} b$ states that entity a stands in relation r to entity b. All labels are arbitrary, so any structure must be learned from the facts themselves.

The model places each entity a at a learned position $\mathbf{x}_a \in \mathbb{R}^d$ and represents each relation r by a learned displacement $\mathbf{r} \in \mathbb{R}^d$, shared by every fact that uses r. Learning $a \xrightarrow{r} b$ means making this displacement carry a to b, so training minimizes the squared norm of each fact's error,

$$\boldsymbol{\epsilon}_{a,r,b} = \mathbf{x}_a + \mathbf{r} - \mathbf{x}_b. \tag{1}$$

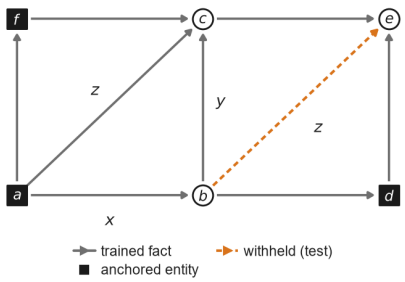


Figure 1: Example relational world.

Composition is then addition of displacements. If the model has learned $a \xrightarrow{x} b$ and $b \xrightarrow{y} c$, following both relations from a gives $\mathbf{x}_a + \mathbf{r}_x + \mathbf{r}_y \approx \mathbf{x}_c$. Our worlds contain a third relation z that joins the two ends of every such path, so one z step must cover the same displacement as the two-step path: $\mathbf{r}_z = \mathbf{r}_x + \mathbf{r}_y$. This compositional law holds in the world by construction but is not built into the model, which must recover it from the facts. Each law spans many entities; training presents only some of the facts that hold, and each experiment withholds others as tests. Because facts constrain only differences between positions, the whole configuration could otherwise shift freely. Anchor penalties fix it by drawing a few entities toward fixed, distinct coordinates, while leaving them learnable.

We collect all positions and displacements as rows of an embedding matrix $E \in \mathbb{R}^{p \times d}$, one row per entity or relation token. The shallow model ($N = 1$) learns E directly, as a lookup table. A depth-N model instead factorizes it as $E = W_N \cdots W_1$, with one $p \times d$ factor and $N - 1$ factors of size $d \times d$. This changes its learning dynamics without changing representational capacity. We use $d = 16$ and $N = 1, 2, 3$, and train by stochastic gradient descent (SGD) on individual facts and anchors, reshuffled each epoch. Appendix B.2 gives initialization, learning rates, and training durations.

3.2 Experimental manipulations

Experiment 1: law-specific emergence. Networks learn 16 compositional laws supported by lattices of different sizes and fact-presentation frequencies (Fig. 2a). Each lattice contains observed and withheld z-facts. We test whether differences in the supporting evidence predict when each law's withheld compositions become retrievable.

Experiment 2: structural identifiability. Observing all three relations need not determine their composition. When observed z-facts connect entity pairs separate from the observed x-then-y paths, the facts leave the law $\mathbf{r}_z = \mathbf{r}_x + \mathbf{r}_y$ undetermined (Fig. 2b). After pretraining, we add one z-fact connecting the endpoints of an observed path and test whether other withheld compositions become retrievable. A comparison structure contains such closing facts from the outset.

Experiment 3: relational integration. Networks first learn two lattices with different entities but shared relation vectors (Fig. 2c). Only the first lattice is anchored, leaving their relative position undetermined. We then connect them with one fact using an existing relation and test inference between other cross-lattice entity pairs.

In Experiments 2 and 3, the *linking fact* is one additional observed relation, trained with the same error objective as the original facts. Linked and no-link conditions begin from identical pretrained weights. The linked condition revisits the added fact each epoch alongside the original facts; the no-link condition continues with the original facts alone. Appendix B.1 specifies the constructions, and Appendix B.2 gives the intervention schedules.

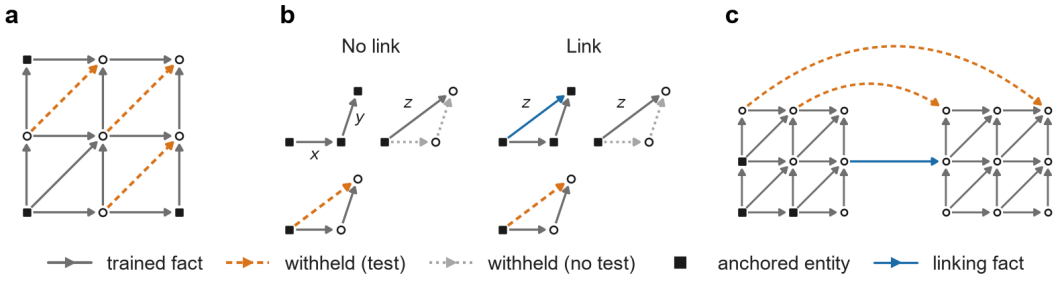


Figure 2: Experimental manipulations of compositional learning.

3.3 Measuring compositional learning

We evaluate inference on facts and comparisons withheld from training. In Experiments 1 and 2, a query $a \xrightarrow{z} b$ places the model's prediction at $\mathbf{q} = \mathbf{x}_a + \mathbf{r}_z$. In Experiment 3, the query applies supplied numbers of x and y steps from a source entity to predict a target in the other lattice (see Appendix B.3).

Compositional accuracy measures how often the correct target is the nearest candidate entity to the query in Euclidean distance. Candidates belong to the relevant lattice or block in Experiments 1 and 2, and to both lattices in Experiment 3. *Geometric error* measures the mean query-to-target distance relative to the current entity spacing in the target's structure. This captures continuous changes that may precede a change in retrieval accuracy. Appendix B.3 gives the query sets and precise metric definitions.

Theoretical predictions start from the same initial weights as the networks and use the same evaluation queries, without fitting their observed trajectories. We summarize results across 200 worlds per depth, treating the world as the unit of replication. Appendix B.4 defines emergence criteria and summary statistics. Separate retention tests track retrieval of previously trained facts after each intervention (Appendix B.5).

4 A Spectral Theory of Compositional Learning

We ask which compositional laws the training evidence determines and how their retrieval develops during learning. The analysis gives a closed-form trajectory for shallow networks and coupled dynamics, integrated numerically, for deeper networks. Appendix A provides the derivations.

4.1 Representing the evidence

Each training fact imposes a linear constraint on the embedding matrix E. For a fact $a \xrightarrow{r} b$, we construct a row of a matrix A with coefficients $+1$ on the source entity and relation, -1 on the target entity, and zero elsewhere. Multiplying this row by E gives the fact's error $\boldsymbol{\epsilon}_{a,r,b}$ (Equation 1). An anchor row has a single $+1$ on the anchored entity, with its fixed coordinates stored in a target matrix C. Stacking the fact and anchor constraints gives the error matrix $AE - C$ and objective

$$\mathcal{L}(E) = \frac{1}{2|D|}\|AE - C\|_F^2, \tag{2}$$

where D is the multiset of training constraints. Repeated rows encode presentation frequency, with linking facts adding one new row. Appendix A.1 gives the full construction.

Two matrices determine the loss gradient, $\nabla_E \mathcal{L} = HE - B$:

$$H = \frac{A^\top A}{|D|}, \qquad B = \frac{A^\top C}{|D|}. \tag{3}$$

Here, H captures the structure and weighting of the constraints, while B incorporates the anchor targets. Since H is symmetric and positive semidefinite, it admits the eigendecomposition $H = Q\Lambda Q^\top$, where the columns q_k of Q are orthonormal eigenvectors and Λ contains their nonnegative eigenvalues λ_k.

Each *spectral mode* $q_k \in \mathbb{R}^p$ specifies a combination of entity and relation embeddings, $q_k^\top E$. Its eigenvalue measures how strongly the training evidence constrains that combination. Positive eigenvalues identify constrained modes while zero eigenvalues identify changes to the embeddings that leave every training error unchanged. This distinction separates what the evidence determines from what remains free.

4.2 Which laws does the evidence determine?

For a law j, such as $\mathbf{r}_z = \mathbf{r}_x + \mathbf{r}_y$, define the contrast $\ell_j \in \mathbb{R}^p$ with coefficients $+1$ on relations x and y, -1 on z, and zero elsewhere. The quantity $\ell_j^\top E = \mathbf{r}_x + \mathbf{r}_y - \mathbf{r}_z$ then measures the law's error. We call the law identifiable if this error is zero in every embedding matrix satisfying the training constraint $AE = C$.

Let Q_+ contain the eigenvectors of H with positive eigenvalues. These modes span the subspace constrained by the evidence, and $Q_+ Q_+^\top$ is the projection onto that subspace.

Proposition 4.1 *If at least one solution of $AE = C$ satisfies the law, the law is identifiable if and only if*

$$\ell_j = Q_+ Q_+^\top \ell_j. \tag{4}$$

This applies the classical estimability criterion for linear contrasts (Searle, 1971). The contrast must lie entirely within the constrained subspace; otherwise, embeddings can fit every training fact while disagreeing about the composition. We quantify this ambiguity in Experiment 2 by

$$\rho_j = \frac{\|\ell_j - Q_+ Q_+^\top \ell_j\|_2}{\|\ell_j\|_2}, \tag{5}$$

which is zero exactly when the law is identifiable.

A linking fact adds a row to A that constrains a previously free direction. In both experiments it removes the only such direction. In Experiment 2, this makes the law identifiable. In Experiment 3, it fixes the relative position of two lattices, determining many cross-lattice comparisons simultaneously. Appendix A.2 proves Proposition 4.1, and Appendix A.6 shows how a linking fact removes the unconstrained translation between the two lattices.

4.3 When do identifiable laws emerge?

Even an identifiable law takes time to learn. For the shallow model, small-step SGD is approximated by the gradient flow

$$\tau \frac{dE}{dt} = B - HE, \qquad \tau = (\eta |D|)^{-1}, \tag{6}$$

where η is the per-constraint learning rate and t is measured in epochs. Let E^* satisfy $AE^* = C$, and define the embedding error along mode k as

$$\zeta_k(t) = q_k^\top \big(E(t) - E^*\big). \tag{7}$$

Under this flow, constrained modes decay independently at rates λ_k/τ, while unconstrained modes remain fixed (Saxe et al., 2014; 2019). A law's error combines these modes through its loadings $a_{jk} = \ell_j^\top q_k$. By Proposition 4.1, an identifiable law depends only on constrained modes, whose errors decay during learning.

Proposition 4.2 *For an identifiable law j, its error under the shallow gradient flow is*

$$\boldsymbol{\delta}_j(t) = \ell_j^\top E(t) = \sum_{k:\lambda_k>0} a_{jk}\boldsymbol{\zeta}_k(0)e^{-\lambda_k t/\tau}. \tag{8}$$

Laws supported by different modes can therefore emerge at different times within one network, with their trajectories dependent on the decay rates and the initial errors (Fig. 3).

Retrieval also depends on the positions of targets and competing candidates. We therefore reconstruct the full embedding trajectory and apply the measures from Section 3.3, using a discrete mean-update predictor described in Appendix A.4. Appendix A.3 provides the derivation of Proposition 4.2.

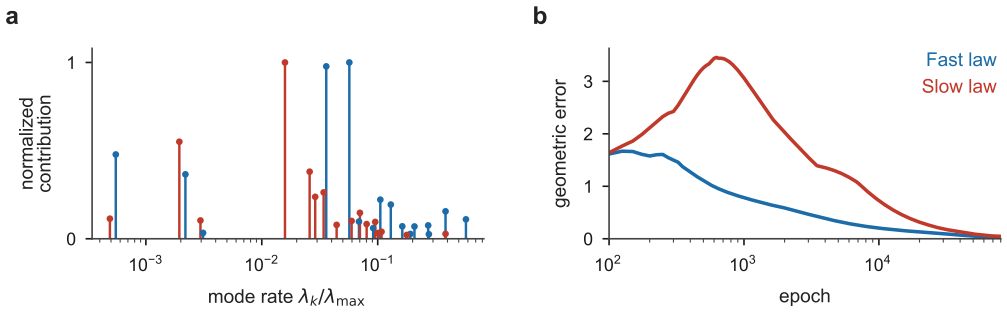


Figure 3: Fast and slow laws differ in spectral support. (a) Contribution of each spectral mode to the initial error of the fast and slow laws, normalized within each law and plotted against its relative decay rate, $\lambda_k/\lambda_{\max}$. (b) Predicted geometric error for the same laws in one $N = 1$ world. The fast law places more weight on faster modes and therefore converges earlier than the slow law.

4.4 How does depth change learning?

The factorization $E = W_N \cdots W_1$ changes how the embeddings are learned without restricting which embeddings the model can represent. Proposition 4.1 therefore applies at every depth.

To describe the factorized dynamics, define the products surrounding each factor W_l as

$$L_l = W_N \cdots W_{l+1}, \qquad R_l = W_{l-1} \cdots W_1, \tag{9}$$

with empty products equal to the identity. Applying the chain rule to $\nabla_E \mathcal{L} = HE - B$ gives the factorized gradient flow (Saxe et al., 2014; Lampinen & Ganguli, 2019),

$$\tau \frac{dW_l}{dt} = L_l^\top (B - HE) R_l^\top. \tag{10}$$

The evolving factors generally couple the spectral modes, so their contributions to a law's error can no longer be predicted independently. We integrate these equations numerically from the same initial weights as the trained network, reconstruct $E(t)$, and apply the retrieval measures from Section 3.3.

Factorization also allows unconstrained modes to change during learning, whereas shallow dynamics preserve their initial values. Thus, depth can affect an underdetermined law through the solution selected by learning, even though the evidence still does not uniquely determine that law. Appendix A.5 derives the factorized dynamics and explains when unconstrained modes can change.

5 Results

5.1 Spectral structure predicts law-specific emergence

Compositional laws supported by different training evidence became retrievable on different timescales. Emergence time t^* was the first evaluation at which held-out compositional accuracy

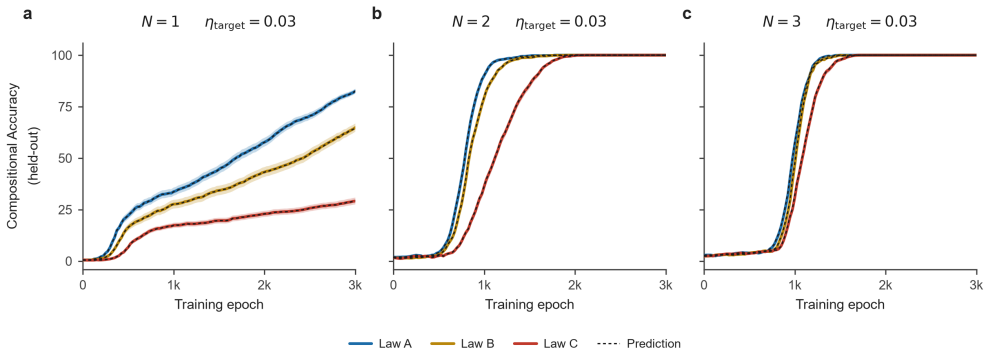


Figure 4: Compositional laws emerge at law-specific times predicted by theory. Three laws are shown. **a–c**, instantaneous accuracy on held-out queries $a \xrightarrow{z} b$. Shading is s.e.m. across 200 worlds. The first 3,000 epochs are shown; Appendix C.1 gives full-run event times. The corresponding geometric trajectories are shown in Figure C.3a–c.

reached 100% and geometric error fell below 0.5, with both sustained for ten consecutive evaluations. Across the 16 laws, median t^* over worlds spanned 4,000–21,438 epochs at $N = 1$, 1,025–2,062 at $N = 2$, and 1,125–1,400 at $N = 3$, a 5.4-, 2.0-, and 1.2-fold range. The corresponding behavioural trajectories are shown in Figure 4.

Starting from each network's initial weights, the theory predicted these differences in timing without fitting the observed trajectories. All 3,200 law–world pairs reached the emergence criterion in both the networks and their predictions at each depth, within 80,000 epochs at $N = 1$ and 10,000 at $N = 2, 3$. Median absolute timing error was zero epochs at every depth, and 100.0%, 99.9%, and 98.8% of predictions fell within one 25-epoch evaluation interval. Predicted and observed $\log_{10}$ emergence times agreed with R^2 = 1.0000, 0.9979, and 0.9793. Agreement extended to the full accuracy and geometric-error trajectories (Table C.1).

5.2 Resolving ambiguity enables compositional generalization

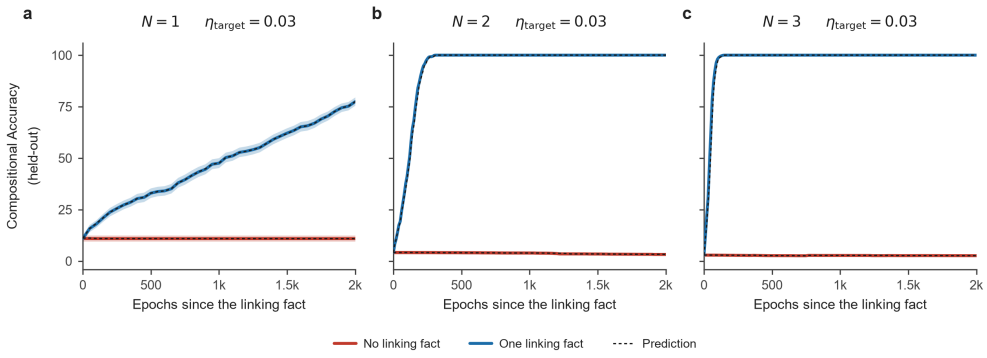


Figure 5: A single linking fact makes an underdetermined composition learnable. **a–c**, held-out accuracy. Shading is s.e.m. across 200 worlds. The corresponding geometric trajectories are shown in Figure C.3d–f.

When the existing evidence left a law ambiguous, adding a single linking fact enabled generalization to withheld compositions. Before the intervention, the law's ambiguity was $\rho = 0.2582$ in every world and depth; the linking fact reduced it to at most 1.4×10^{-15}. Two thousand epochs later, linked networks retrieved 77.4%, 100.0%, and 100.0% of held-out compositions at $N = 1, 2, 3$, compared with 10.9%, 3.2%, and 2.6% in matched no-link controls continuing from the same pre-trained weights (Figure 5). Every linked world reached 90% accuracy, with median recovery times

of 1,850, 160, and 60 epochs. Linked geometric error fell 1.82-, 29.7-, and 36.3-fold over the same period, whereas without the link accuracy changed by less than one percentage point and geometric error by at most 4% (Figure C.3d–f).

The predicted dynamics captured how quickly compositional accuracy improved after the ambiguity was resolved. The median absolute error in recovery time was 0, 10, and 10 epochs, no more than one evaluation interval (50, 10, and 10 epochs), and 100.0%, 100.0%, and 99.5% of predictions fell within one interval. The R^2 of $\log_{10}$ recovery time was 0.9986, 0.9860, and 0.8735. The lower R^2 at $N = 3$ reflects reduced relative timing precision despite a median absolute error of only 10 epochs; Appendix C.4 examines this difference and reports agreement over the full accuracy and geometric-error trajectories.

To test whether the intervention disrupted earlier knowledge, we replayed 30 worlds at each depth under both structural assignments and both intervention conditions. All 72 original training facts, including the 64 premises, remained correct at every sampled evaluation over the full post-intervention horizons (Appendix B.5). Thus, the new compositions became retrievable without any detected loss of earlier knowledge on these evaluation grids.

5.3 One linking fact unlocks cross-structure inference

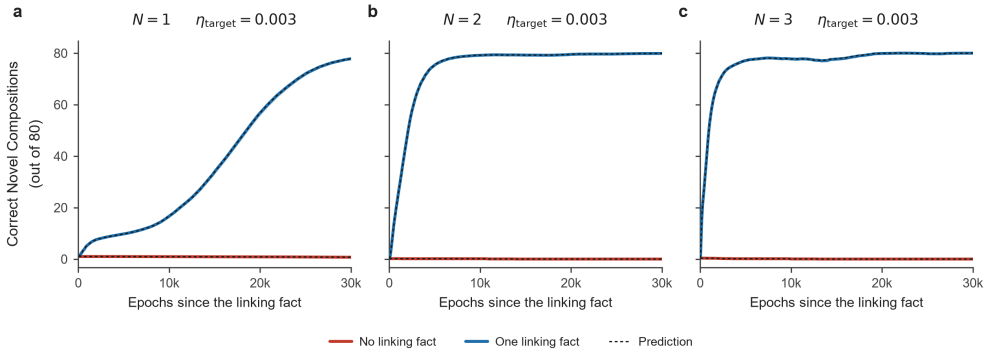


Figure 6: One linking fact unlocks many cross-structure inferences. **a–c**, absolute number of correctly answered held-out cross-structure comparisons, averaged across worlds. Shading is s.e.m. across 200 worlds. The corresponding geometric trajectories are shown in Figure C.3g–i.

A linking fact also enabled generalization between two structures that had been learned separately. Before the link, networks answered almost none of the 80 held-out cross-structure comparisons: 1.0, 0.2, and 0.4 at $N = 1, 2, 3$. Thirty thousand epochs later, linked networks answered 77.8, 79.9, and 80.0, compared with 0.7, 0.0, and 0.0 in matched no-link controls continuing from the same pretrained weights (Figure 6). Median time to 72 correct comparisons was 24,082, 3,771, and 2,250 epochs. Cross-structure geometric error fell 7.73-, 21.0-, and 163-fold with the link and changed by at most 5% without it (Figure C.3g–i).

The same dynamics predicted how long these cross-structure inferences took to become retrievable. The median absolute error in the time to reach 72 correct comparisons was zero epochs at every depth, and 99.5%, 100.0%, and 95.0% of predictions fell within the observed crossing interval, the gap between the last evaluation below criterion and the first at or above it (median 943, 155, and 58.5 epochs; Appendix B.4). The R^2 of $\log_{10}$ unlock time was 0.9996, 0.9994, and 0.9996. The predictions also tracked the accuracy and geometric-error trajectories (Table C.1).

This integration came with a largely temporary loss of previously learned premises (Figure C.2). At least one of the 24 original premise facts became incorrect in 162, 170, and 136 of 200 linked worlds at depths 1–3; none of the no-link worlds showed such a loss. Of these 468 affected world–depth cases, 467 regained full premise retrieval at subsequent checks through the end of the run. One depth-1 world had not recovered by 40,000 epochs. The predicted and observed mean retention curves differed by at most 0.17 percentage points on the shared evaluation grids. Thus, the theory captured both the gain in cross-structure inference and the accompanying disruption of earlier knowledge.

6 Discussion and Limitations

We asked how training evidence determined which compositional laws a neural network could recover and when they became retrievable. In controlled additive worlds, our theory connected the constraints imposed by observed facts to the learning dynamics of relational embeddings. For identifiable laws, the theory predicted both their retrieval trajectories and emergence times across shallow and deep networks. When a law remained underdetermined, the theory identified the source of the ambiguity and predicted how retrieval developed after a linking fact resolved it. Linking two previously separate relational structures likewise enabled many cross-structure inferences, whose development was predicted by the same dynamics. Together, these findings show how the organization of training evidence can account for differences in compositional learning, from laws acquired at different times to coordinated gains in generalization after new evidence is introduced.

Our results distinguished two reasons why compositional retrieval could remain poor despite continued training. An identifiable law could remain poorly retrieved because the modes supporting it were still being learned. An underdetermined law instead remained compatible with conflicting solutions, even when every training constraint was satisfied. This distinction matters when interpreting held-out performance. A split may test whether a learner recovers a consequence of the observations, or whether its inductive bias selects an answer that the observations leave unresolved. In our setting, the identifiability criterion established which problem the learner faced before its trajectory was examined. It therefore provided a way to separate insufficient evidence from slow learning.

Our theory was developed in synthetic worlds with known additive laws and predefined relation types. This allowed us to isolate compositional learning, but left open whether the same principles apply to heterogeneous or non-additive relations. The analysis was also limited to shallow and deep linear networks. Although factorization introduced nonlinear learning dynamics, the experiments did not establish whether our predictions extend to nonlinear architectures. Future work should test whether the spectral structure of relational evidence remains informative in these settings, including on knowledge-graph benchmarks such as FB15k-237 (Toutanova & Chen, 2015) and WN18RR (Dettmers et al., 2018). Matched relational tasks with human learners or nonlinear networks (e.g. large language models) could also provide a controlled test by comparing the empirical order of compositional emergence with our predictions. Such tests would clarify which aspects of our account generalize beyond the additive models studied here.

AI USE STATEMENT

Generative AI tools were used to help develop and refine the theoretical framing, formulate mathematical claims, design experiments, implement and debug analysis code, interpret results, search the literature, structure the manuscript, and edit prose. The author inspected the mathematical derivations, tested the code and numerical predictions against saved outputs, checked cited work, and reviewed all AI-assisted text. The author takes responsibility for the final content of this work, including claims, code, text, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses synthetic relational worlds and computational models. It involves no human participants, personal data, or animal subjects. We identify no direct ethical risks beyond the general obligation to report theoretical assumptions and computational results accurately.

REPRODUCIBILITY STATEMENT

The relational objective, theoretical assumptions, prediction equations, evaluation measures, and experimental manipulations are specified in Sections 3–4. Further derivations and procedural details appear in Appendices A–B.

REFERENCES

Timothy E. J. Behrens, Timothy H. Muller, James C. R. Whittington, Shirley Mark, Alon B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018. doi: 10.1016/j.neuron.2018.10.002.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644. URL https://aclanthology.org/2024.tacl-1.16/.

Ronald B. Dekker, Fabian Otto, and Christopher Summerfield. Curriculum learning for human compositional generalization. *Proceedings of the National Academy of Sciences*, 119(41): e2205582119, 2022. doi: 10.1073/pnas.2205582119.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2D knowledge graph embeddings. In *AAAI Conference on Artificial Intelligence*, 2018. URL https://arxiv.org/abs/1707.01476.

Steven M. Frankland and Joshua D. Greene. Two ways to build a thought: Distinct forms of compositional semantic representation across brain regions. *Cerebral Cortex*, 30(6):3838–3855, 2020. doi: 10.1093/cercor/bhaa001.

Kelvin Guu, John Miller, and Percy Liang. Traversing knowledge graphs in vector space. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 318–327, 2015. doi: 10.18653/v1/D15-1038. URL https://aclanthology.org/D15-1038/.

Graeme S. Halford, William H. Wilson, and Steven Phillips. Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, 14(11):497–505, 2010. doi: 10.1016/j.tics.2010.08.005.

Dharshan Kumaran and James L. McClelland. Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, 119(3):573–616, 2012. doi: 10.1037/a0028681.

Andrew K. Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1809.10374.

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.

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. URL https://arxiv.org/abs/2301.05217.

Stephanie Nelli, Lukas Braun, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Neural knowledge assembly in humans and neural networks. *Neuron*, 111(9):1504–1516.e9, 2023. doi: 10.1016/j.neuron.2023.02.014.

Core Francisco Park, Maya Okawa, Andrew Lee, Hidenori Tanaka, and Ekdeep Singh Lubana. Emergence of hidden capabilities: Exploring learning dynamics in concept space. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.19370.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6120.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116 (23):11537–11546, 2019. doi: 10.1073/pnas.1820226116.

Simon Schug, Seijin Kobayashi, Yassir Akram, Maciej Wołczyk, Alexandra Proca, Johannes von Oswald, Razvan Pascanu, João Sacramento, and Angelika Steger. Discovering modular solutions that generalize compositionally. In *International Conference on Learning Representations*, 2024.

Shayle R. Searle. *Linear Models*. John Wiley & Sons, New York, 1971.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1902.10197.

Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 57–66, 2015. doi: 10.18653/v1/W15-4007. URL https://aclanthology.org/W15-4007/.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2307.05596.

Yongyi Yang, Core Francisco Park, Ekdeep Singh Lubana, Maya Okawa, Wei Hu, and Hidenori Tanaka. Swing-by dynamics in concept learning and compositional generalization. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.08309.

Dagmar Zeithamova and Alison R. Preston. Flexible memories: Differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. *The Journal of Neuroscience*, 30(44): 14676–14684, 2010. doi: 10.1523/JNEUROSCI.3250-10.2010.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702, 2023. doi: 10.18653/v1/2023.emnlp-main.971. URL https://aclanthology.org/2023.emnlp-main.971/.

A MATHEMATICAL DERIVATIONS

We give the training-system construction, proofs of Propositions 4.1 and 4.2, and the extensions needed for finite learning rates, depth, and linking interventions. The algebraic identifiability result is exact. The trajectory results are exact for the stated mean dynamics, which approximate the per-row SGD used in the experiments.

A.1 CONSTRUCTING THE TRAINING SYSTEM

Let $e_u \in \mathbb{R}^p$ select token u. Each row of E stores the embedding of one entity or relation. A fact $a \xrightarrow{r} b$ contributes the row $(e_a + e_r - e_b)^\top$ to A and a zero row to C. An anchor for entity a with target $c_a \in \mathbb{R}^d$ and weight w_a contributes $\sqrt{w_a} e_a^\top$ to A and $\sqrt{w_a} c_a^\top$ to C. Thus Equation 2 is the sum of squared fact errors and weighted squared anchor errors, normalized by twice the number of rows. We use $w_a = 1$. Anchors penalize deviations from a target; they do not clamp an embedding. Repeated facts contribute repeated rows, preserving their presentation frequencies.

The matrices have dimensions $A \in \mathbb{R}^{|D| \times p}$, $C \in \mathbb{R}^{|D| \times d}$, $H \in \mathbb{R}^{p \times p}$, and $B \in \mathbb{R}^{p \times d}$. Since $H = A^\top A/|D|$ is positive semidefinite, its positive-eigenvalue subspace is the row space of A. Write the orthogonal projectors onto this subspace and its complement as

$$P_+ = Q_+ Q_+^\top, \qquad P_0 = I - P_+. \tag{11}$$

The complement is $\ker A$, the directions unobserved by training. Our generated systems are consistent: their constructed coordinates provide a solution to $AE = C$.

A.2 IDENTIFIABILITY: PROOF OF PROPOSITION 4.1

Fix a consistent solution $E^\star$ that satisfies the law, $\ell_j^\top E^\star = 0$. Every other solution is $E = E^\star + V$ with $AV = 0$. If $P_0 \ell_j = 0$, then $\ell_j^\top V = 0$ for every such V, so all solutions satisfy the law.

Conversely, suppose $v = P_0 \ell_j \neq 0$. For any nonzero $u \in \mathbb{R}^d$, set $V = vu^\top$. Then $AV = 0$, while

$$\ell_j^\top (E^\star + V) = \|v\|_2^2 u^\top \neq 0. \tag{12}$$

This gives a solution that violates the law. Hence identifiability is equivalent to $P_0 \ell_j = 0$, or Equation 4.

We measure the unresolved fraction of a nonzero contrast by

$$\rho_j = \frac{\|P_0 \ell_j\|_2}{\|\ell_j\|_2}. \tag{13}$$

Thus $\rho_j = 0$ exactly when the law is identifiable. The same argument applies to any linear query contrast, provided a consistent solution gives its intended value. It characterizes the set of solutions, independently of the optimizer. An optimizer may select a particular solution when $\rho_j > 0$, and nearest-target retrieval may be correct even with a nonzero geometric residual.

A.3 SHALLOW DYNAMICS: PROOF OF PROPOSITION 4.2

Differentiating the objective gives

$$\nabla_E \mathcal{L} = \frac{1}{|D|} A^\top (AE - C) = HE - B. \tag{14}$$

To first order in the per-row step size η, one epoch of SGD sums these row gradients at the epoch's starting state. Its update is $-\eta A^\top (AE - C)$. Taking the continuous-time approximation in epoch units gives $\tau \dot{E} = B - HE$, with $\tau = (\eta |D|)^{-1}$. Finite-step, order-dependent corrections are omitted by this approximation.

For any consistent $E^\star$, $HE^\star = B$. Projecting the embedding error onto an eigenvector q_k gives the row vector $\boldsymbol{\zeta}_k(t) = q_k^\top (E(t) - E^\star) \in \mathbb{R}^{1 \times d}$, with

$$\tau \dot{\boldsymbol{\zeta}}_k = -\lambda_k \boldsymbol{\zeta}_k, \qquad \boldsymbol{\zeta}_k(t) = \boldsymbol{\zeta}_k(0) e^{-\lambda_k t/\tau}. \tag{15}$$

Positive modes decay independently, and zero modes remain at initialization. Although $E^\star$ need not be unique, its positive-mode projections are fixed by the evidence:

$$q_k^\top E^\star = \lambda_k^{-1} q_k^\top B \quad (\lambda_k > 0). \tag{16}$$

Thus the positive-mode errors used in Proposition 4.2 require no choice of an otherwise unconstrained solution.

For an identifiable law, $\ell_j = P_+ \ell_j$ and $\ell_j^\top E^\star = 0$. Expanding in the eigenbasis yields

$$\begin{aligned} \boldsymbol{\delta}_j(t) &= \ell_j^\top (E(t) - E^\star) \\ &= \sum_{k:\lambda_k>0} (\ell_j^\top q_k) \boldsymbol{\zeta}_k(t) = \sum_{k:\lambda_k>0} a_{jk} \boldsymbol{\zeta}_k(0) e^{-\lambda_k t/\tau}, \end{aligned} \tag{17}$$

which proves the proposition. For a non-identifiable law, choosing a consistent $E^\star$ that satisfies the law leaves the additional term $\ell_j^\top P_0 (E(0) - E^\star)$, which persists under shallow learning. Several modes can contribute with different signs and rates, so the norm of the law residual need not decay monotonically.

The full continuous-time prediction is

$$\widehat{E}(t) = E^\star + P_0 (E(0) - E^\star) + \sum_{k:\lambda_k>0} q_k \boldsymbol{\zeta}_k(0) e^{-\lambda_k t/\tau}. \tag{18}$$

Applying each evaluation query to this matrix retains the evolving source, relation, target, and distractor embeddings. The law residual alone does not specify a nearest-target decision. Predicted event times use the same sampled evaluations and thresholds as the network, without fitting a rate or threshold to its trajectory.

A.4 Discrete shallow predictions and their relation to SGD

The implemented shallow predictor uses the discrete mean update

$$\widehat{E}_{t+1} = \widehat{E}_t - \eta A^\top (A \widehat{E}_t - C). \tag{19}$$

Writing $\sigma_k = |D| \lambda_k$ for the eigenvalues of $A^\top A$, its closed form is

$$\widehat{E}_t = E^\star + Q \operatorname{diag}\big((1 - \eta\sigma_k)^t\big) Q^\top (E(0) - E^\star). \tag{20}$$

The code uses the minimum-norm least-squares solution for $E^\star$. All positive-mode errors decay when $0 < \eta < 2/\sigma_{\max}$; zero-mode components remain unchanged. For small $\eta\sigma_k$, the decay factor approximates $e^{-\eta\sigma_k t} = e^{-\lambda_k t/\tau}$.

Equation 20 is exact for Equation 19, not for the ordered product of per-row SGD updates. Per-row SGD also preserves the shallow null component exactly, since each update lies in the row space of A, but its constrained trajectory can differ at finite step size. The experiments compare these predicted and observed trajectories directly.

A.5 Factorized depth and unconstrained directions

For $E = W_N \cdots W_1$, define

$$L_l = W_N \cdots W_{l+1}, \qquad R_l = W_{l-1} \cdots W_1, \tag{21}$$

with empty products $L_N = I_{p\times p}$ and $R_1 = I_{d\times d}$. A variation in layer W_l gives $dE = L_l \, (dW_l) \, R_l$. The chain rule therefore gives

$$\nabla_{W_l} \mathcal{L} = L_l^\top (HE - B) R_l^\top. \tag{22}$$

Gradient flow yields Equation 10. Because the layer products evolve, the environmental modes generally interact. We integrate the layer equations numerically and reconstruct $\widehat{E}(t; N)$ at each evaluation point; Appendix B.2 gives the integration settings.

The induced embedding dynamics are

$$\tau \dot{E} = \sum_{l=1}^{N} L_l L_l^\top (B - HE) R_l^\top R_l. \tag{23}$$

Although $P_0(B - HE) = 0$, in general $P_0 L_l L_l^\top (B - HE) \neq 0$. The embedding's null component can consequently evolve. More specifically, $P_0 \dot{W}_N = 0$, but $P_0 E = (P_0 W_N) W_{N-1} \cdots W_1$ can change through the remaining factors. If $P_0 W_N(0) = 0$, the end-to-end null component instead remains zero.

For example, take $A = (1, 0)$, $C = 1$, $W_2 = (1, 1)^\top$, $W_1 = 2$, and $\tau = 1$. Then $B - HE = (-1, 0)^\top$, $\dot{W}_2 = (-2, 0)^\top$, and $\dot{W}_1 = -1$, giving $\dot{E} = (-5, -1)^\top$. The unconstrained second entry changes, whereas shallow flow leaves it fixed. This is optimizer-dependent evolution in a direction that the evidence does not identify. Proposition 4.1 continues to hold; in our parameterization, factorization does not restrict the set of representable embedding matrices.

A.6 How a linking fact changes the system

A new fact contributes one row $g^\top = e_a^\top + e_r^\top - e_b^\top$, with zero target. With no new tokens or anchors, the post-intervention system is

$$A' = \begin{pmatrix} A \\ g^\top \end{pmatrix}, \qquad C' = \begin{pmatrix} C \\ 0 \end{pmatrix}, \qquad A'^\top A' = A^\top A + g g^\top. \tag{24}$$

Its null space is $\ker A' = \ker A \cap \ker g^\top$. A new row can therefore remove an unconstrained direction only if it has a component along that direction. Its effect is shared by all query contrasts that depend on the removed freedom.

In Experiment 3, the first lattice's anchors and within-lattice facts fix its geometry and the shared offsets. The second lattice's internal facts fix its shape but leave its translation free. If v has entries one on the second lattice's entities and zero elsewhere, this freedom is $E \mapsto E + v u^\top$ for arbitrary $u \in \mathbb{R}^d$, and $\ker A = \mathrm{span}(v)$. The linking row runs from the first lattice to the second, so $g^\top v = -1$ and removes this freedom. Each cross-lattice query has error contrast $c = e_a + n_x e_x + n_y e_y - e_b$, also with $c^\top v = -1$. Hence the link determines all these contrasts, not just its own trained pair. This is a statement about solutions of the constraints; the time needed to retrieve the targets still depends on learning dynamics.

For post-intervention predictions, we recompute $H' = A'^\top A'/(|D|+1)$ and $B' = A'^\top C'/(|D|+1)$. The per-row learning rate is retained, so the epoch time constant becomes $\tau' = (\eta(|D|+1))^{-1}$. Each predicted condition continues from its own predicted pre-intervention state. The network conditions similarly copy their common trained state. Predictions are never reset to the trained network's intervention state.

B EXPERIMENTAL DETAILS

B.1 SYNTHETIC RELATIONAL WORLDS

Worlds contained arbitrary entity and relation labels. The primary construction instantiated additive laws $\mathbf{r}_z = \mathbf{r}_x + \mathbf{r}_y$ through facts of the form $a \xrightarrow{x} b$, $b \xrightarrow{y} c$, and $a \xrightarrow{z} c$. Each law appeared across multiple entity sets. Some valid z-facts were included in training and distinct valid z-facts were withheld as compositional queries. This separates retrieval of an unseen composite fact from memorization of a trained triple.

In the integration experiment, two relational lattices used the same relation types but separate entity sets. Before intervention, all within-lattice constraints could be learned while the lattices' relative translation remained unconstrained. One cross-lattice fact removed that freedom in the linked condition. The matched no-link condition continued from the identical network state without receiving the resolving constraint.

Experiment 1: law-specific emergence. The world uses eight lattice specifications (m, n, r_x, r_y), each instantiated for two laws with distinct relation tokens: $(5, 4, 1, 1)$, $(4, 4, 2, 2)$, $(4, 4, 3, 4)$, $(4, 4, 4, 6)$, $(5, 4, 2, 8)$, $(4, 4, 1, 4)$, $(4, 5, 4, 8)$, and $(4, 4, 3, 6)$. Here m, n are the numbers of entity positions along the two axes, indexed by $0 \le i < m$ and $0 \le j < n$. The values r_x, r_y count presentations of each corresponding edge per epoch; trained diagonal edges use the same multiplicity as the second-axis edges. The $(m-1)(n-1)$ diagonal cells correspond to composite facts $(i, j) \xrightarrow{z} (i+1, j+1)$. The constructor samples 25% without replacement for training, rounded to an integer and with at least one included. Each lattice has unit-weight anchors at $(0, 0)$, $(m-1, 0)$, and $(0, n-1)$, using its generated target coordinates: 48 anchors across the 16 lattices. The world specifications were selected during exploratory development to separate emergence times; they are not a random sample of relational topologies. The three displayed laws are `L3_1`, `L2_1`, and `L5_0`, selected once from pooled results and then held fixed; all 16 laws enter the reported statistics.

Experiment 2: structural identifiability. Each world has two blocks, each with 16 premise triads and four composite facts. In the closed block, the four composite facts connect the endpoints of the first four observed premise triads. In the open block, they instead connect four additional entity pairs sharing no entities with any premise triad; the premise triads themselves remain connected two-edge paths. Each block anchors the initial entity of every premise triad and the middle and final entities of its first triad (18 anchors per block). The open block also anchors the head of each additional pair, leaving its tail unanchored: 40 unit-weight anchors in total, all at generated target coordinates. The intervention adds the composite fact between the endpoints of the first open-block premise triad; no anchor is added. Both assignments of open and closed blocks are run for every world. The evaluation set excludes facts present after the intervention and is held fixed across conditions.

Experiment 3: relational integration. Each world uses two 3×3 lattices sharing offsets $x, y, z = x + y$, with three anchor targets in the first lattice and none in the second. The added fact connects the lattices using the existing x relation. The offsets and initial lattice origins are Gaussian draws in 16 dimensions. The second lattice is then translated so that its origin lies one x step beyond the first lattice's final first-axis entity; this makes the linking fact true in the constructed world. The first lattice's anchors are at $(0, 0)$, $(1, 0)$, and $(0, 1)$.

B.2 TRAINING AND NUMERICAL PREDICTION

The model minimized Equation 2 using SGD with a random permutation of all fact and anchor rows in every epoch, a fixed per-row learning rate, and no adaptive optimizer. Shallow networks learned E directly. Deeper networks used $E = W_N \cdots W_1$, with $W_N \in \mathbb{R}^{p \times d}$ and the other layers in $\mathbb{R}^{d \times d}$. Here $d = 16$, so the factorization does not restrict the class of embedding matrices. All initial entries were Gaussian with zero mean and standard deviation 0.01 for shallow embeddings and 0.05 for each deep layer.

The selected worlds have indices $s = 0, \ldots, 199$, with model initialization seed $1000 + s$. World generation uses seed s in Experiments 1 and 3 and $1000 + s$ in Experiment 2. Thus the Experiment 2 world generator and model initializer use the same seed in separate generator instances; their random

Table B.1: Run settings for the selected 200 worlds per depth. One epoch visits every constraint row once, including repeated facts and anchors. For staged runs, $\eta = \eta_{\text{target}}/\lambda_{\max}(A_0^\top A_0)$ is set using the pre-intervention matrix A_0 and retained after the intervention. Shallow predictions use the discrete closed form. The original intervals are supplemented by denser replay evaluations for Experiment 3 and the retention controls (Appendix B.5).

Experiment	Setting	$N = 1$	$N = 2$	$N = 3$
1	η_{target}	0.03	0.03	0.03
	Total epochs	80,000	10,000	10,000
	Original evaluation interval	25	25	25
	RK4 substeps per epoch	—	4	8
2	η_{target}	0.03	0.03	0.03
	Pre-intervention epochs	50,000	6,000	3,500
	Post-intervention epochs	40,000	8,000	4,000
	Original evaluation interval	50	10	10
	RK4 substeps per epoch	—	4	8
3	η_{target}	0.003	0.003	0.003
	Pre-intervention epochs	200,000	65,000	55,000
	Post-intervention epochs	40,000	40,000	40,000
	Original evaluation interval	2,500	500	250
	RK4 substeps per epoch	—	1	1

streams are not independent within a world. Experiment 1 uses presentation-order seed $2000 + s$. Experiments 2 and 3 use order seed 7 before intervention and 8 after intervention, shared across worlds; their uncertainty summaries therefore vary world draws and initializations, not order seeds independently.

Shallow predictions use Equation 20; deeper predictions use fourth-order Runge–Kutta integration of Equation 10, with the depth-specific substeps in Table B.1. Predictions start from the same initial weights as training and use the same fixed query set. At intervention, the two network conditions copy the network's pre-intervention state, while the two predicted conditions continue from their own predicted pre-intervention state. The resolving fact is one additional distinct training constraint, revisited each subsequent epoch alongside the original constraints; this is not a single-update experiment. Pretraining durations were chosen separately by depth using settling diagnostics, and post-intervention trajectories are indexed by epochs since the added fact.

The normalized learning rate is $\eta_{\text{target}} = \eta\lambda_{\max}(A^\top A)$. Positive modes are selected using an eigenvalue cutoff of 10^{-9} times the largest eigenvalue; contrasts with $\rho < 10^{-7}$ are classified as identifiable. The observed ambiguity values are separated by many orders of magnitude from this cutoff. The candidate-spacing denominator includes 10^{-12} to avoid division by zero.

B.3 Held-out queries and evaluation measures

In the emergence and identifiability experiments, each held-out query was an unseen composite fact $a \xrightarrow{z} b$, evaluated at

$$\mathbf{q} = \mathbf{x}_a + \mathbf{r}_z. \tag{25}$$

Candidates were restricted to entities in the corresponding lattice (Experiment 1) or block (Experiment 2). Compositional accuracy was the proportion of targets ranked first by Euclidean distance. Geometric error was computed with Equation 27.

In the integration experiment, held-out items were cross-structure comparisons. If source a and target b were separated by n_x steps along relation x and n_y steps along relation y, the query was

$$\mathbf{q} = \mathbf{x}_a + n_x\mathbf{r}_x + n_y\mathbf{r}_y. \tag{26}$$

The integers n_x, n_y are supplied query coefficients derived from the constructed world's coordinates. The task tests their application to learned offsets, rather than discovery of the path. Retrieval uses all 18 entities across the two lattices. Of the 81 cross-lattice pairs, the directly trained linking pair is excluded, leaving 80 queries. Geometric error uses Equation 27, with spacing computed among the

nine destination entities. Thus the retrieval pool includes distractors from the source lattice while the geometric reference scale remains that of the destination lattice.

For M held-out queries, geometric error is

$$\mathcal{E}_{\text{geo}} = \frac{1}{M} \sum_{i=1}^{M} \frac{\|\mathbf{q}_i - \mathbf{x}_{b_i}\|_2}{s_i + 10^{-12}}, \tag{27}$$

where s_i is the median nearest-neighbour distance among the learned entity embeddings in query i's reference structure, recomputed at each evaluation. The reference structure is the retrieval candidate set in Experiments 1 and 2 and the nine-entity destination lattice in Experiment 3. The denominator expresses error in units of the target structure's current spacing; 10^{-12} prevents division by zero. Retrieval is counted as correct if no candidate is strictly closer than the target, matching the implementation's treatment of ties.

Full-pool performance before linking was low on average but not zero in every world. At intervention, 177, 198, and 198 of 200 worlds answered none of the 80 queries correctly at depths 1–3. The largest initial scores were 9, 27, and 63, respectively. Without a link, depth-1 world 1 retained its score of 9 through 30,000 epochs; the depth-2 and depth-3 maxima both occurred in world 91 and fell to zero by that time. We report raw counts and matched no-link controls rather than subtracting a baseline.

B.4 Emergence criteria and summary statistics

For Experiment 1, t^* was the first evaluation at which held-out accuracy equalled 100% and normalized geometric error was below 0.5, with both criteria sustained for ten consecutive evaluations. Experiment 2 recovery was the first evaluation reaching 90% held-out accuracy. Experiment 3 unlock was the first evaluation with at least 72 of 80 comparisons correct among all 18 candidates. The latter two criteria do not require a subsequent period of sustained success.

Timing agreement is $R^2 = 1 - \sum_i (y_i - \widehat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$, where $y_i = \log_{10} t_i$ and only pairs of finite, positive event times are included. This measures agreement without fitting a regression; it is not squared correlation. Events attained at intervention contribute to attainment counts and absolute errors but have no logarithmic time. This excludes one depth-1 world from Experiment 2's log-time comparison. All Experiment 3 event times are positive.

Experiments 1 and 2 use the original evaluation intervals in Table B.1. Experiment 3 uses the denser replay grid described in Appendix B.5, with predictions evaluated at every network evaluation time. The observed crossing interval is the gap from the last check below criterion to the first meeting it. The reported within-interval share counts worlds in which the absolute difference between observed and predicted first-success times is no greater than that world's observed crossing interval. This is a tolerance on sampled event times, not a test of exact continuous-time coincidence.

At depths 1–3, median Experiment 3 crossing intervals are 943, 155, and 58.5 epochs, with maxima 1,779, 424, and 250. Median absolute timing error is zero at each depth; the 95th percentiles are 0, 144.8, and 202.5 epochs. The within-interval shares are 99.5%, 100.0%, and 95.0%. These finite-resolution measures accompany the log-time R^2 in Section 5.3.

Trajectory R^2 uses raw accuracy or $\log_{10}$ geometric error, computed within each world (and law in Experiment 1) before averaging; constant trajectories with undefined R^2 are omitted. Each evaluation carries equal weight. Experiment 3 trajectory agreement uses the shared replay grid over the first 30,000 post-intervention epochs; its denser early sampling gives that part of the trajectory more weight. Experiment 1 uses the full depth-specific runs, and Experiment 2 uses the first 2,000 post-intervention epochs for trajectory agreement and effects. Experiment 3 effects are reported at 30,000 epochs; unlock times use the full 40,000-epoch follow-up. Worlds are the replication unit, with the two Experiment 2 structural assignments averaged within each world before inference. Reported confidence intervals are 95% Student-t intervals over the 200 world-level values; fold-change intervals are computed in log space and exponentiated. Accuracy means are arithmetic; geometric curves use geometric means and log-space standard errors.

B.5 Replay validation and retention controls

We replayed all 200 Experiment 3 worlds at each depth from their original seeds, preserving initialization, presentation order, learning rates, and durations. At every original post-intervention evaluation, the replay reproduced the stored held-out hits and geometric scores exactly. It additionally scored all 18 retrieval candidates and the original training facts. Theory predictions reproduced the stored predicted hits exactly and geometric scores to relative tolerance 10^{-9}.

The replay grid combines the original evaluation times with checks every ten epochs through epoch 500, every 50 through epoch 2,000, and a geometrically spaced sequence thereafter with factor 1.05, rounded to integer epochs. Over 40,000 post-intervention epochs this gives 158, 218, and 294 evaluations at depths 1–3. Predictions include every network evaluation time; finer prediction-only points are used to draw the curves but not to compare event times. Pre-intervention training lasts 200,000, 65,000, and 55,000 epochs, respectively.

Retention tests retrieval of an original fact using candidates in its own structure. In Experiment 3 we score both the 24 original x/y premises and all 32 original training facts, excluding the newly added link. All are correctly retrieved at intervention. A distinct loss counts a fact that fails at any sampled evaluation; a simultaneous loss counts facts failing at the same evaluation. These differ when different facts fail at different times.

At least one premise fails in 162, 170, and 136 linked worlds at depths 1–3, and in no no-link world. Among affected worlds, the median numbers of distinct premise losses are 1, 2, and 2, with maxima 4, 10, and 14. Maximum simultaneous premise losses are 3, 10, and 10. Scoring all 32 original facts gives the same affected-world counts, but distinct-loss medians of 1, 3, and 2 and maxima of 5, 11, and 16. The maximum simultaneous losses across all original facts are 4, 11, and 11. Of 468 affected world–depth cases, 467 return to full retrieval at all subsequent evaluations through 40,000 epochs; depth-1 world 22 has not recovered by that horizon. These observations do not exclude losses between checks or later failures.

The theory predicts premise losses in 162, 171, and 135 worlds. Its affected set matches the network at depth 1, adds world 144 at depth 2, and misses world 10 at depth 3. The minima of the network's mean premise-retention curves are 95.50%, 93.48%, and 95.67%; the corresponding theoretical minima on the shared grids are 95.48%, 93.46%, and 95.56%. These minima describe retention, not prediction accuracy. The maximum pointwise gaps between the mean curves are 0.021, 0.104, and 0.167 percentage points.

For Experiment 2, we replayed seeds 0–29 at each depth with both assignments of the open and closed blocks. Each of the 180 cells included linked and no-link continuations. All 64 original premises and all 72 original training facts remained correct at every evaluation in both conditions. The depth-specific post-intervention horizons were 40,000, 8,000, and 4,000 epochs, sampled at 899, 824, and 412 evaluations. Every original evaluation within those horizons was reproduced exactly. The result is no detected loss in this sample and on these grids, rather than a guarantee of stability between evaluations.

C SUPPLEMENTARY RESULTS

This appendix collects full-run emergence times (Appendix C.1), premise retention (Appendix C.2), geometric trajectories (Appendix C.3), and prediction agreement (Appendix C.4). Figures and tables are numbered by appendix, so each can be cited directly.

C.1 FULL-HORIZON EMERGENCE TIMES

Figure C.1 shows every law-world's event time and the distribution of emergence times across worlds for each law. Unlike the early trajectories in Figure 4, these summaries include the full 80,000-epoch shallow run and 10,000-epoch deep runs. All 3,200 law-worlds reach criterion at each depth, so none are omitted for censoring. A fixed law label fixes its lattice dimensions and presentation counts, not its sampled diagonal constraints, target vectors, or initial projections. These draws alter the environmental spectrum, multimode coefficients, and retrieval geometry. The per-world predictor incorporates them together; the present experiment does not isolate their individual contributions to timing variance.

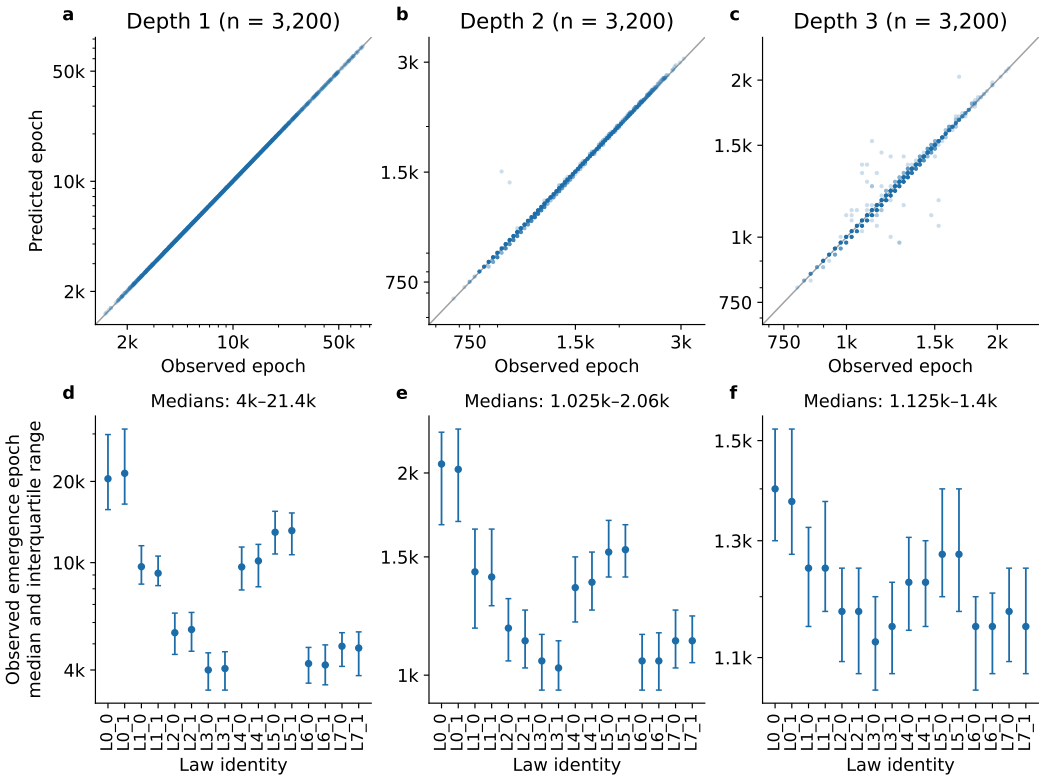


Figure C.1: Emergence times across all 16 laws and 200 worlds. **a–c**, predicted versus observed emergence epochs for all 3,200 law-worlds at each depth; the diagonal denotes equality. Points overlap when event times coincide on the evaluation grid. **d–f**, median observed event time for each fixed law, with interquartile ranges across worlds. All timing axes are logarithmic; limits differ across depths. Error bars describe between-world variation, not uncertainty in the median.

C.2 PREMISE RETENTION DURING INTEGRATION

Figure C.2 shows the transient premise-retrieval losses described in Section 5.3. The sample, candidate sets, and recovery criteria are specified in Appendix B.5.

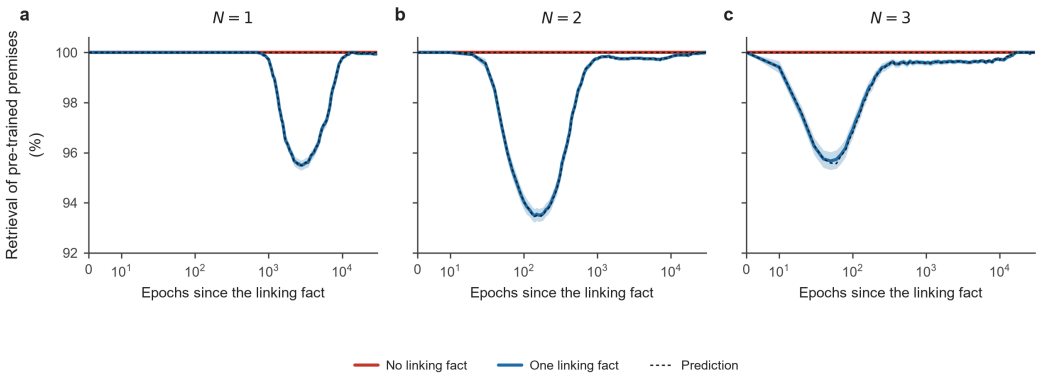


Figure C.2: Linking two structures can temporarily disrupt retrieval of original premises. Mean retrieval of the 24 original x/y premise facts in Experiment 3, scored within each fact's own lattice, over the same 200 worlds per depth as Figure 6. The ordinate refers to these premises, not all 32 original training facts. Blue and red curves show linked and no-link networks; dashed curves show theory predictions without trajectory-specific fitting. Shading is s.e.m. across worlds. Time is linear below ten epochs and logarithmic above. The first 30,000 post-intervention epochs are shown; recovery assessments use the full 40,000-epoch run. Curves connect sampled evaluations. Mean retention can approach 100% while an individual world still has an error.

C.3 GEOMETRIC TRAJECTORIES

Figure C.3 shows the geometric error behind Figures 4–6, one row per experiment. Each row uses the same worlds, runs and predictions as the corresponding main-text figure, with the geometric error defined in Section 3.3.

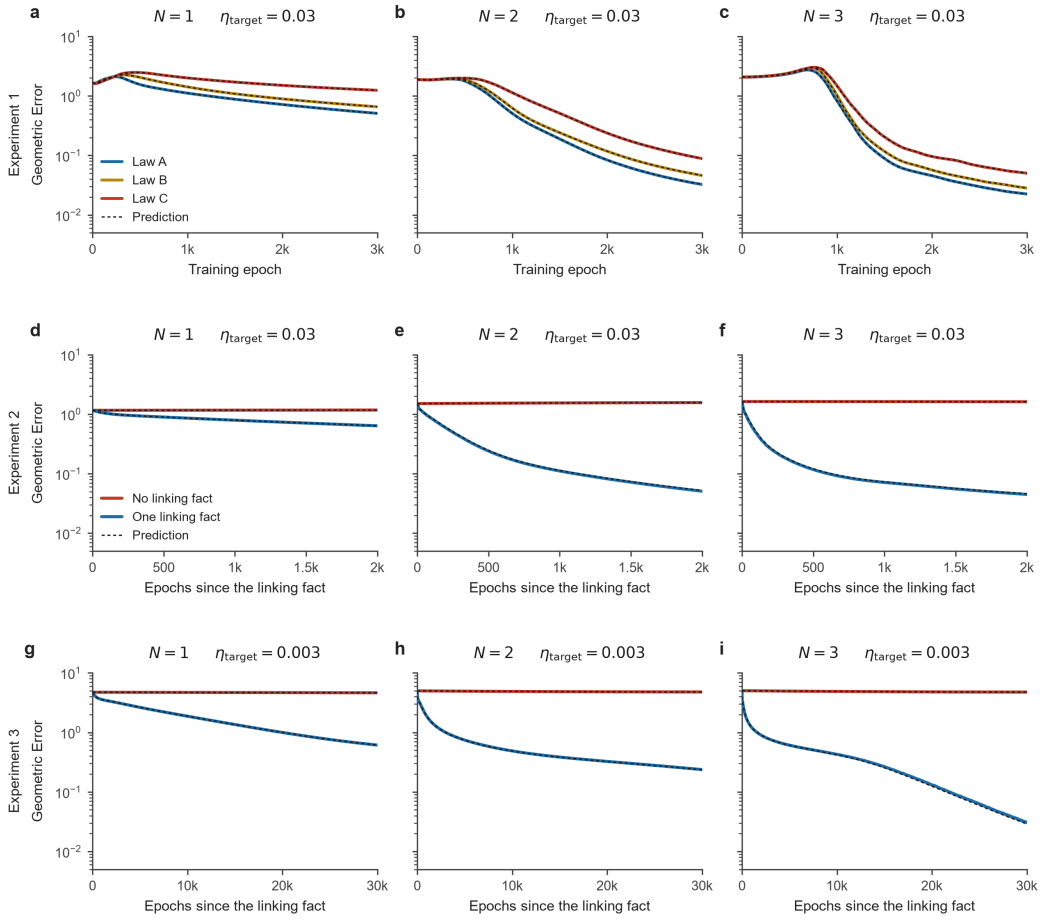


Figure C.3: Geometric error in the three experiments. **a–c**, Experiment 1, and **d–f**, Experiment 2: query-to-target distance normalized by median nearest-neighbour spacing among candidates. **g–i**, Experiment 3: distance between each cross-structure query and its correct target, normalized by median nearest-neighbour spacing among the nine destination entities. Curves are geometric means across 200 worlds, shaded by log-space s.e.m.; dashed lines show predictions.

C.4 PREDICTION AGREEMENT

Table C.1 compares predicted and observed trajectories over each experiment's full horizon, using the per-trajectory R^2 defined in Appendix B.4.

Table C.1: Agreement between predicted and observed trajectories. Values are means over worlds (over law-worlds in Experiment 1) of the per-trajectory R^2, and for Experiment 2 also medians over worlds of the mean absolute accuracy difference and of the geometric-error ratio. Experiments 2 and 3 report the linked condition; no-link trajectories are nearly flat, and at the 95th percentile their predictions differ by at most 0.17 percentage points in accuracy and a factor of 1.002 in geometric error. Horizons are 80,000 epochs at $N = 1$ and 10,000 at $N = 2, 3$ in Experiment 1, and 2,000 and 30,000 epochs after the linking fact in Experiments 2 and 3.

Experiment	Measure	$N = 1$	$N = 2$	$N = 3$
1	Accuracy, mean R^2	0.9999	0.9997	0.9989
	Log geometric error, mean R^2	1.0000	1.0000	0.9999
2	Accuracy, mean R^2	0.9978	0.9808	0.9359
	Accuracy, median absolute error (pp)	0.00	0.22	0.25
	Log geometric error, mean R^2	1.0000	0.9994	0.9977
	Geometric error, median ratio	1.000	1.020	1.027
3	Correct comparisons, mean R^2	1.0000	0.9999	0.9992
	Log geometric error, mean R^2	1.0000	1.0000	0.9993

Two thousand epochs after the intervention in Experiment 2, linked accuracy exceeded no-link accuracy by 66.5, 96.8, and 97.4 percentage points at $N = 1, 2, 3$ (95% CIs 61.7–71.3, 95.3–98.3, and 96.0–98.7). Thirty thousand epochs after the link in Experiment 3, linked networks answered 77.8, 79.9, and 80.0 of 80 comparisons at $N = 1, 2, 3$ (95% CIs 77.2–78.5, 79.8–80.0, and 80.0–80.0), with all 80 correct in 129, 196, and 200 of 200 worlds, and geometric error fell 7.73-, 21.0-, and 163-fold (95% CIs 7.46–8.02, 19.9–22.2, and 137–194).

In Experiment 2, the recovery-time R^2 is lower at $N = 3$ (0.8735) than at $N = 2$ (0.9860). This is not solely a low-variance artifact: from $N = 2$ to $N = 3$ the observed log-time variance falls from 0.0311 to 0.0267, while the mean squared log-time error rises from 0.000435 to 0.00337. Absolute agreement on the evaluation grid and relative timing precision therefore give complementary assessments. At $N = 1$, one world already above 90% accuracy at the intervention is excluded from R^2.