GEOMETRY OF COGNITIVE MAPS IN MINIMAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

An effective cognitive map organizes states into a compact coordinate system whose geometry—the relative arrangement of states—preserves task-relevant relations and makes them explicit for inference and planning. Yet determining how task structure constrains this geometry is difficult in neural population recordings and artificial neural networks, where representations are obscured by noise, finite sampling, learning dynamics, and architectural choices. Here, we model cognitive maps as regularized hidden-layer activations in linear neural networks trained to predict the next state. Building on previously derived theoretical representations, we prove that their dimensionality can be analytically computed as the number of connected components plus the number of globally compatible independent actions, minus one. Applying this theory across spatial and non-spatial tasks, our results reveal an organizing principle of cognitive maps not captured by standard state-transition models: how transitions are grouped into relational operations could shape cognitive-map geometry. The theory also provides several insights over neural representations, including how shared action encoding supports structural transfer and how increasing similarity among action representation (i.e. action merging) may serve as a mechanism for acquiring a low-dimensional cognitive map. Cognitive-map representations therefore reflect an agent's internal definitions of states and relational operations rather than the external environment alone. By translating candidate task models into quantitative predictions of neural population geometry, our theory may also open a principled route to inferring how an agent internally represents a cognitive map.

1 INTRODUCTION

Cognitive maps organize states, places, events, or concepts by their relations, providing a common representational framework for spatial and non-spatial knowledge (Tolman, 1948; Garvert et al., 2017; Whittington et al., 2022; Igarashi et al., 2022; Courellis et al., 2024; Whittington et al., 2020). By encoding how states are connected, cognitive maps support planning (Pfeiffer & Foster, 2013; Stachenfeld et al., 2017; Nyberg et al., 2022; Stöckl et al., 2024) and the transfer of relational structure across environments (Mark et al., 2020).

One of the central functions of cognitive maps is relational compression: for a given task, they preserve relations among states that are relevant for behavior while reducing other variation, thereby expressing those relations in a compact coordinate system (Nieh et al., 2021; Behrens et al., 2018). Within this system, relations such as distance and direction become explicit, supporting novel inference (Constantinescu et al., 2016; Garvert et al., 2017; Park et al., 2020). Therefore, we argue that the coordinates that make explicit the relations preserved through this compression define the representation geometry of a cognitive map.

Determining how the task itself constrains the geometry of a cognitive map is especially challenging in neural population data. Biological and observational noise, finite sampling, task confounding factors can obscure the representation geometry expressed in recorded population activity and bias dimensionality estimates (Altan et al., 2021; Chun et al., 2026; Recanatesi et al., 2019; Gao et al., 2017; Rigotti et al., 2013). Learning further complicates this inference because representation geometry changes as task structure is acquired (Niv, 2019; Wójcik et al., 2026). Together, these factors

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.

make it difficult to identify which components of an observed representation geometry are dictated by a given task structure.

As an alternative, artificial neural networks (ANNs) are used as model systems in which input and output can be precisely controlled and the resulting internal representations directly examined. Yet this approach also faces an identifiability problem: as overparameterized systems, ANNs can implement the same function with different internal geometries, with network architecture, initialization, and optimization dynamics determining which solution is selected (Frei et al., 2023; Mehrer et al., 2020; Braun et al., 2025; Flesch et al., 2022; Farrell et al., 2023). For example, recurrent networks can develop place- and grid-cell-like codes during navigation (Cueva & Wei, 2018; Banino et al., 2018), but their emergence depends sharply on target encoding and other design choices (Schaeffer et al., 2022). Successful training alone therefore cannot identify which aspects of a learned representation express cognitive-map geometry.

State-space models are also used to model cognitive maps such as methods based on successor representations (SR) (Dayan, 1993; Stachenfeld et al., 2017; Yoo et al., 2024; Carvalho et al., 2024; Kahn et al., 2025) or hidden Markov model (HMM) (Mark et al., 2020; George et al., 2021). Although they have yielded important insights into predictive and transferable structure, the geometries they induce or infer depends on the policy chosen and some prior assumptions about the latent structure and latent variables. Moreover, these methods primarily characterize predictive occupancy or latent transition structure rather than deriving how task structure constrains the geometry of representations learned within a neural network. Consequently, they provide a less direct account of how neural representational geometry emerges during the acquisition of a cognitive map.

Here, we develop a theory of cognitive-map representation geometry in feedforward neural networks performing next step prediction from samples (Recanatesi et al., 2021; Chen et al., 2024; Levenstein et al., 2024). This provides a network implementation of representing relational graphs from sampled transitions. Building on the analytic minimum-representation-norm solutions of Braun et al. (2025), we formally link this representational geometry to task requirement (Theorem 2.4). We further prove that its dimensionality is determined by the number of connected graph components and globally compatible independent actions (Theorem 2.5).

This theory offers two advantages for studying cognitive maps. First, it separates the aspects of ANN representation geometry determined by task structure from variation introduced by network initialization and optimization dynamics. Second, like state-space modeling approaches, it imposes generic state and action codes, allowing the same formulation to be applied across spatial and non-spatial tasks. Together, these features provide an analytically tractable normative framework for formulating precise, testable hypotheses about neural population geometry, including questions that are difficult to address using existing methods such as the successor representation (SR).

We first apply the theory to spatial navigation and family-relation tasks. In both cases, it yields representations whose leading principal component axis aligns with task-relevant variables. We then use a series of tasks to investigate distinct aspects of this cognitive-map geometry. Disconnected-graph tasks reveal the conditions that support structural transfer and reuse. A random-walk task based on the graph of Garvert et al. (2017) examines how merging similar relations gives rise to low-dimensional representations. A composed navigation–family-relation task tests how multiple task demands can be accommodated within a unified representation. Finally, analyses of nonlinear networks and neuron subsampling assess the generalizability of the theory.

2 Theoretical Results

We study cognitive maps through the hidden representation in feedforward neural networks performing next state prediction based on current state and action. Building on the analytic representation solutions provided in Braun et al. (2025), we establish a formal relationship between the resulting representation geometry, particularly their principal-component subspace, and a vector space spanned by action-consistent task coordinates. We further derive an exact decomposition of such representation dimensionality into two contributions: number of disconnected graph components and globally compatible independent actions.

2.1 COGNITIVE MAP AS MRNS IN PREDICTIVE NETWORKS

Consider a finite deterministic Markov Decision Process (MDP) with state space $\mathcal{S} = \{1, 2, \ldots, D_s\}$, action space $\mathcal{A} = \{1, 2, \ldots, D_a\}$, and deterministic transition function $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$. Let $e_j^{(D)} \in \mathbb{R}^D$ denote the jth standard basis vector, whose jth entry is 1 and whose remaining entries are 0.

Suppose that we observe a finite dataset of P transitions $\mathcal{D} = \{(s_i, a_i, s_i' = T(s_i, a_i))\}_{i=1}^P$. These transitions define the task structure as an *action-labeled* directed graph: the states in $\mathcal{S}$ are its vertices, and each observed transition (s_i, a_i, s_i') defines a directed edge from s_i to s_i' labeled by a_i. For transition i, the network input is the concatenation of the one-hot encodings of the current state and action, and the prediction target is the one-hot encoding of the next state:

$$x_i = \left[e_{s_i}^{(D_s)}; e_{a_i}^{(D_a)}\right], \qquad y_i = e_{s_i'}^{(D_s)}.$$

Stacking the P transitions columnwise gives the input and output matrices:

$$X = \begin{bmatrix} S \\ A \end{bmatrix} = [x_1, \ldots, x_P], \quad Y = [y_1, \ldots, y_P].$$

In this paper, we study cognitive map representation in a two-layer linear network with hidden width D_h, input weights $W_1 \in \mathbb{R}^{D_h \times (D_s + D_a)}$, and readout weights $W_2 \in \mathbb{R}^{D_s \times D_h}$. Its prediction and hidden representation are

$$\widehat{Y} = W_2 W_1 X, \qquad H = W_1 X.$$

We rephrase Theorem 3.5 in Braun et al. (2025), which gives a complete characterization of the minimum-representation-norm solutions defined below.

Definition 2.1. A minimum-representation-norm solution (MRNS) is any pair (W_1, W_2) solving

$$\min_{W_1, W_2} \|W_1 X\|_F^2 + \|W_2\|_F^2 \quad \text{s.t.} \quad W_2 W_1 \Sigma_{xx} = \Sigma_{yx}, \qquad \Sigma_{xx} = \frac{1}{P} X X^\top, \quad \Sigma_{yx} = \frac{1}{P} Y X^\top.$$

Theorem 2.2. *Let $Z := Y X^+ X \in \mathbb{R}^{D_s \times P}$ have compact singular value decomposition*

$$Z = M N O^\top,$$

where $M \in \mathbb{R}^{D_s \times r}, O \in \mathbb{R}^{P \times r}, M^\top M = O^\top O = I_r, N = \mathrm{diag}(n_1, \ldots, n_r)$ and $r = \mathrm{rank}(Z)$.

Assume that $D_h \geq r$, then a pair (W_1, W_2) is a minimum-representation-norm solution (MRNS) if and only if it can be written as

$$W_1 = R\sqrt{N} O^\top X^+ + \Gamma\big(I - X X^+\big), \qquad W_2 = M\sqrt{N} R^\top,$$

for some $R \in \mathbb{R}^{D_h \times r}, R^\top R = I_r$, and arbitrary $\Gamma \in \mathbb{R}^{D_h \times (D_s + D_a)}$. The corresponding hidden representation is

$$H := W_1 X = R\sqrt{N} O^\top,$$

which is independent of Γ. Thus, all MRNS hidden representations differ only by the choice of the semi-orthogonal embedding R. Consequently, every MRNS training representation is an isometric embedding of the same intrinsic representation $\sqrt{N}\, O^\top$.

Consequently, whenever training reaches a global optimum of the objective in Definition 2.1, the learned representation necessarily belongs to this analytically characterized solution family, offering a way to study network representations independently of training dynamics. Importantly, although individual hidden units are not identifiable, the PCA subspace geometry in this representation is uniquely determined by the task structure given by Z (Appendix A). We next show that the dominant PCA subspace of the representation is spanned by action-consistent task coordinates.

2.2 ACTION-CONSISTENT TASK COORDINATES DETERMINE MRNS GEOMETRY

A task coordinate assigns a real number to each state. We represent such a coordinate by a mapping $f : \mathcal{S} \to \mathbb{R}$, whose value $f(s)$ is the coordinate value of state s. Along an observed transition (s, a, s'), the coordinate changes by $f(s') - f(s)$. We call the coordinate *action-consistent* if there

exists an action-increment mapping $\delta : \mathcal{A} \to \mathbb{R}$ such that $f(s') - f(s) = \delta(a)$ for every observed transition. Here $\delta(a)$ is the scalar displacement along this coordinate assigned to action a, shared by all transitions carrying the same action label. For example, in planar navigation the horizontal coordinate has fixed increments under east, west, and stay actions, independently of the current location.

We collect all action-consistent task coordinates in the following vector space (see Appendix B for a proof of $\mathcal{F}$ being a vector space in $\mathbb{R}^{D_s}$):

$$\mathcal{F} := \left\{ f : \mathcal{S} \to \mathbb{R} \;\middle|\; \exists\, \delta : \mathcal{A} \to \mathbb{R}, \quad f(s') - f(s) = \delta(a) \quad \forall\, (s, a, s') \in \{(s_i, a_i, s_i')\}_{i=1}^{P} \right\}$$
$$\cong \left\{ \mathbf{f} \in \mathbb{R}^{D_s} \;\middle|\; \exists\, \boldsymbol{\delta} \in \mathbb{R}^{D_a}, \quad (Y^\top - S^\top)\mathbf{f} = A^\top \boldsymbol{\delta} \right\}.$$

For the second definition, we collected the values of f, δ in its corresponding vector $\boldsymbol{f} \in \mathbb{R}^{D_s}, \boldsymbol{\delta} \in \mathbb{R}^{D_a}$ and equivalently $f(s) = e_s^T \mathbf{f}, \delta(a) = e_a^T \boldsymbol{\delta}$. In such expression, the basis of $\mathcal{F}$ are task coordinates compatible with the task structure. Next, we want to understand the dimensionality of this vector space and its relationship to the PCA space of the MRNS representation of cognitive map (i.e. MRNS geometry).

Assumption 1. $YY^\top = n_{\text{vis}} I_{D_s}$. That is to say, every state occurs exactly n_{vis} times among the next-state targets.

Assumption 2. For every state, the observed transitions include a self-transition labelled by the same stay action a_0. This assumption can always be enforced by augmenting the sampled transitions with a stay transition at each state.

Definition 2.3. Let $H \in \mathbb{R}^{D_h \times p}$ denote an MRNS representation, and let $C_p = I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p^\top$ be the centering matrix. If $\lambda_{\max}$ denotes the theoretically maximal eigenvalue of the PCA covariance $\frac{1}{p}HC_pH^\top$, we define the task-determined representation dimensionality as

$$d_{\text{task}} := \dim \ker\left(\frac{1}{p}HC_pH^\top - \lambda_{\max} I_{D_h}\right).$$

Thus, d_{task} is the geometric multiplicity of the theoretically maximal eigenvalue in the PCA spectrum of an MRNS representation. It can also be interpreted as number of principal components whose eigenvalue equals the maximal theoretical value. Here d_{task} is defined at the level of the network representation.

Theorem 2.4. *Under Assumptions 1 and 2, the task-determined subspace $\mathcal{F}$ coincides with the left singular subspace of the task-dependent matrix Z associated with its largest singular value $\sigma_{\max}$.*

For every MRNS, $\sigma_{\max}$ is the theoretical upper bound of the PCA spectrum of its hidden representation, and the eigenspace attaining this bound is determined exactly by $\mathcal{F} \cap \mathbf{1}_{D_s}^\perp$, where $\mathbf{1}_{D_s}^\perp$ denotes the subspace orthogonal to the all-ones state vector. Therefore,

$$d_{\text{task}} = \dim\big(\mathcal{F} \cap \mathbf{1}_{D_s}^\perp\big) = \dim \mathcal{F} - 1.$$

A proof is given in Appendix B.1.

Thus, although an MRNS representation is defined in the hidden-unit space and is identifiable only up to a semi-orthogonal transformation, its dominant PCA subspace is determined entirely by the underlying task structure: it is spanned by the nonconstant action-consistent task coordinates. Consequently, d_{task} counts the number of linearly independent, nonconstant action-consistent task coordinates represented in $\mathcal{F}$.

2.3 Computing task dimensionality from action labelled transition graph

Whereas Section 2.2 identifies equivalence of MRNS dominant PCA subspace to the span of nonconstant action-consistent task coordinates, we next show that task dimensionality admits a complementary decomposition that reveals the contributions of graph connectivity and globally compatible independent action increments.

Assumption 3. Every action labels at least one observed transition.

Let G be the number of connected components of the state-transition graph when edge directions are ignored. Under Assumption 3, each $f \in \mathcal{F}$ determines a unique action-increment vector δ (see Appendix B.2). Let d_{inc} denote the dimension of the resulting vector space of globally compatible independent increment vectors δ.

Theorem 2.5. *Under Assumptions 1– 3, task dimensionality decomposes as*

$$d_{\text{task}} = G - 1 + d_{\text{inc}}.$$

If the number of actions consist of the stay action and p inverse-action pairs (i.e. $D_a = 2p+1$), and every observed transition has its oppositely directed inverse transition with the paired label, then

$$0 \le d_{\text{inc}} \le p, \qquad d_{\text{task}} \le G + p - 1.$$

Thus p is the maximum possible number of independent actions, whereas d_{inc} counts those that remain independent after enforcing consistency across all observed paths (i.e., requiring all sequences of observed transitions connecting the same pair of states to yield the same net increment). Path relations and action incompatibility can make this contribution strictly smaller than p (see examples in Appendix B.4). Dimensionality is therefore not determined by the number of states, actions, or apparent environmental coordinates alone, but by how action labels constrain task coordinates across the entire cognitive map. We next apply these theoretical results across spatial and non-spatial tasks.

3 Numerical Results

3.1 The dominant PCA subspace captures interpretable task variables

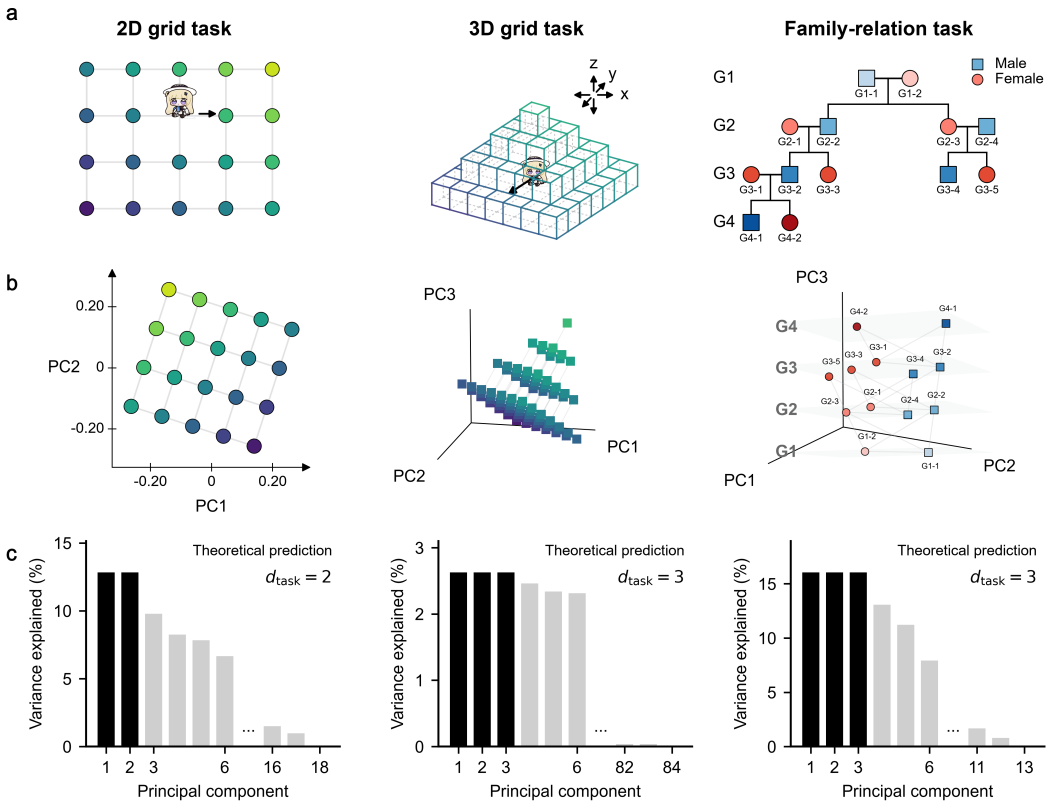


Figure 1: **Dominant PCA subspaces of 2D/3D grid navigation and family-relation tasks capture interpretable task variables.** **a**, Schematics of the 2D grid, 3D grid, and family-relation tasks. **b**, Dominant PCA subspaces of the corresponding MRNS representations. **c**, Explained-variance spectra; black bars mark the numerically identified dominant components.

We tested whether the task dimensionality d_{task} predicted directly from task structure through Theorem 2.5 matched that obtained from the corresponding MRNS, and examined what these dimensions encoded. We considered three benchmark tasks: a 2D grid, a 3D grid, and a family-relation task (Figure 1a). The two spatial tasks used a stay action and bidirectional movement along each

spatial axis, and the non-spatial family-relation task used stay, husband/wife, and eight directed parent–child actions distinguishing fathers from mothers and sons from daughters. For each task, we constructed transition samples under Assumption 1, analytically obtained an MRNS, and applied PCA to the centered sample representations.

In the navigation tasks, d_{inc} has a direct spatial interpretation: it counts the movement directions whose increments can vary independently while remaining consistent across every path through the grid. In the 2D grid, horizontal and vertical movements provide two such directions. Opposite actions have increments of equal magnitude and opposite sign, while horizontal increments are independent from vertical ones. Thus, $d_{\text{inc}} = 2$. The 3D grid adds an independent increment along the third spatial axis, giving $d_{\text{inc}} = 3$. The family-relation task shows why counting action pairs alone can overestimate dimensionality. Besides stay, its husband/wife labels and eight directed labels encoding parent and child relations form five inverse-action pairs, which could nominally contribute five independent increments. The pedigree, however, contains alternative paths between the same people. For example, the increment from Father directly to Son must equal the sum of the increments from Father to Mother and from Mother to Son. This and analogous path relations make two of the five candidate increments dependent, leaving $d_{\text{inc}} = 3$ (see Appendix B.4).

Consistent with these predictions, the dominant PCA subspaces recovered the underlying 2D and 3D grid geometries, whereas the family-relation representation clearly distinguished individuals by generation and sex (Figure 1b,c). Across the three tasks, the PCA spectra contained exactly the predicted two, three, and three dominant components, respectively (Figure 1c). Thus, d_{inc} reflects the number of relational directions that remain independently adjustable across the entire task, rather than simply the number of named relations, and provides an intuitive account of the representation geometry recovered by the numerical results.

3.2 Shared action encoding enables cross-environment structural generalization

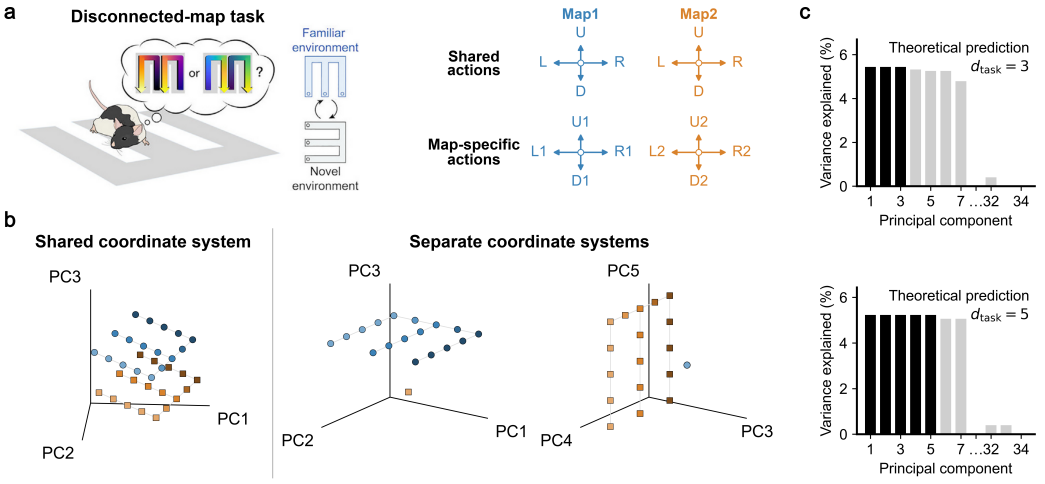


Figure 2: **Shared action encoding enables coordinate reuse in a disconnected-map navigation task. a**, Schematic of the disconnected-map task, adapted with permission from SciDraw.io (CC-BY; DOI: `https://doi.org/10.5281/zenodo.3926011`) and Tang et al. (2023). **b**, Dominant PCA subspaces under shared (left) and map-specific action encoding (right). The five-dimensional map-specific subspace is visualized from two orthogonally rotated views. **c**, Explained-variance spectra for shared (top) and map-specific (bottom) actions.

Across species, relational structure can be reused despite changes in sensory content: rodents aligned prefrontal neural representation across geometrically similar environments despite different sensory stimuli (Tang et al., 2023; Samborska et al., 2022), and human can transfer learned graph structure to new stimulus sets to infer unobserved trajectories (Mark et al., 2020). Yet the mechanism of such transfer remains unresolved: what enables representations in different environments to be identified

as instances of the same abstract relation, and when can they be expressed in a common coordinate system?

Here, we drew on the cross-environment W-maze task of Tang et al. (2023), in which rats transferred a learned spatial rule between visually distinct familiar and novel environments. To investigate the potential mechanism for such cross-environment transfer, we modeled the two environments as disconnected maps with fixed states and within-map transitions and manipulated whether action labels were encoded separately for each map or shared across map (Figure 2a).

The map-specific-action condition assigned the representations of two different environments to orthogonal subspaces (Figure 2b,c). In contrast, the shared-action condition reduced task dimensionality from five to three and the participation ratio (PR) from 26.62 to 25.95, yielding a compressed representation that reused two spatial coordinates across maps while retaining one coordinate for map identity (Figure 2b,c). This is consistent with experimental evidence that prefrontal representations generalized across environments while hippocampal representations preserved their distinction (Tang et al., 2023; Samborska et al., 2022). Thus, structural generalization may arise when corresponding transitions across environments are labelled as the same action, organizing otherwise separate maps within a shared coordinate system.

3.3 Action merging compresses relational knowledge into a cognitive map

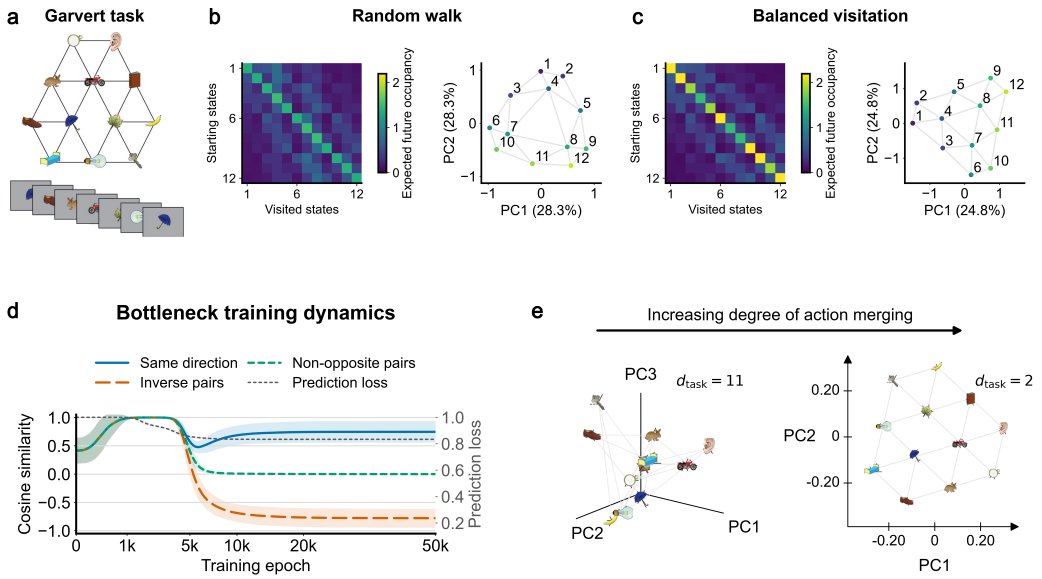


Figure 3: **Action merging serves as a mechanism for acquiring a low-dimensional cognitive map. a**, Schematic of 12-object relational graph adapted from Garvert et al. (2017). **b,c**, Successor-representation matrices and their first two PCs at $\gamma = 0.8$ under random-walk (**b**) and balanced-visitation (**c**) policies. **d**, Training dynamics of the learned two-dimensional linear bottleneck across 100 random seeds; curves show means and bands show one standard deviation. **e**, Dominant PCA subspaces for the edge-specific (left) and fully action-merged (right) tasks.

A knowledge map usually encodes pairwise associations, whereas a compact and efficient cognitive map organizes them to support computation. In the study by Garvert et al. (2017), participants viewed random walks on a hidden 12-object graph (Figure 3a). Despite reporting no awareness of its structure, their response times tracked graph communicability, while hippocampal–entorhinal adaptation tracked graph distance, consistent with a map-like representation. These findings raise a central question: how does learning transform pairwise relationship into an organized cognitive map that supports inference over indirect relations ?

To investigate this question, we first consider the successor representation (SR), a widely studied theory of policy-dependent predictive representations. We computed SRs under two policies on the same graph (Figure 3b,c): a random walk that chose uniformly among neighboring states, and a

balanced-visitation policy that used degree-dependent stay probabilities to equalize target visitation. PCA of both SRs recovered an approximate triangular organization. Across a broader discount-factor sweep, the SR geometry varied with policy and γ (Figure S2). Notably, the balanced-visitation policy produced a more regular geometry than the random walk, suggesting that Assumption 1 helps recover cognitive-map geometry beyond our theory. However, the SR primarily characterizes the predictive representation induced by a given transition policy; it does not by itself explain how learning from pairwise transition samples gives rise to such an organized low-dimensional geometry.

We next applied our theory to the same task, first assigning a distinct action identity to every directed edge. This edge-specific definition provided no reusable relational identity across transitions, yielding a task dimensionality of eleven and a high-dimensional, unstructured representation (Figure 3e, left).

Yet intriguingly, when trained on this task with a narrow hidden-layer bottleneck (Figure S1), a two-layer linear network spontaneously merged the representation of actions corresponding to the same directions in the graph's triangular layout (blue curve), and gradually resembled the organization recovered by the SR (Figure 3d). We therefore hypothesized that this layout captures a candidate organization of the human cognitive map. We then defined *action merging* by assigning a shared action identity to edges pointing in the same direction. Reapplying our theory reduced task dimensionality from eleven to two and PR from 11.00 to 7.70, transforming the unstructured representation into a low-dimensional compact cognitive map (Figure 3e, right). Taken together, these results suggest that action merging may serve as a mechanism for acquiring an efficient cognitive map from transition experience, by recognizing which edges instantiate the same category of relational operation and thereby supporting flexible inference and planning.

The comparison shows that the SR and our theory capture complementary aspects of cognitive-map organization. The SR constructs policy-dependent transition structure. Our theory considers the relational identity of edges—which transitions an agent interprets as instances of the same relation—an aspect rarely examined by models that encode task structure primarily through transition or emission matrices. Moreover, we showed that these relational identities need not be specified *a priori*, but can emerge naturally through learning under a representational bottleneck. Learning which transitions instantiate the same relation may be especially important in non-spatial tasks, where transitions may express qualitatively distinct relations, and may provide a general principle for organizing cognitive maps.

3.4 Factorized states encode multiple relational graphs within a shared representation

Adaptive behavior often simultaneously depends on multiple relational structures. An episodic memory, for example, may encode both where an event occurred and what or whom it involved. Humans can combine separately learned abstract dimensions into a unified cognitive map (Park et al., 2020). Consistent with this, hippocampal representations capture both spatial and non-spatial relationships (Deuker et al., 2016; Buzsáki et al., 2022; Zutshi et al., 2025). How can these distinct relational structures coexist within a shared representation without interfering with one another?

To investigate this question, we constructed a combined task involving both spatial navigation and family relations, and compared two alternative ways of defining its states. In the one-hot superposition scenario, 13 family members were assigned to a 4×5 grid whose 20 locations served as the states (Figure S3a). Navigation and family-relation actions therefore imposed conflicting constraints on the same states, collapsing task dimensionality to zero (Figure S3b,c). In the Cartesian-product scenario, each state paired a location with a person, yielding 260 joint states in total (Figure S3d). The resulting five-dimensional representation comprised two spatial and three family-relation dimensions (Figure S3e,f).

These results recast interference between cognitive maps as a problem of state composition. Factorizing task-relevant variables allows otherwise incompatible relational structures to coexist in orthogonal subspaces of a shared representation. More broadly, this suggests that the brain may accommodate concurrent task demands by representing experience compositionally, rather than as a single indivisible state identity.

3.5 Partial recovery of the task subspace in nonlinear and under-sampled representations

Although linear networks permit precise analysis, we asked whether the predicted geometry persists in nonlinear networks. We compared unregularized networks with networks receiving pre- or post-activation representation-norm regularization, evaluating recovery from the leading task-dimensional PCs (Table S1) and from the best-matching task-dimensional combination among the first 20 PCs (Table S2). Across all tested conditions, the predicted cognitive-map geometry was detectable even without representation-norm regularization. Moreover, in many conditions, the best-matching task-dimensional subspace among the first 20 PCs matched the prediction more closely than the same-dimensional leading-PC subspace, whereas representation-norm regularization generally shifted task-relevant structure toward the leading, higher-variance PCs, reducing its mean PC index from 8.10 to 5.96.

A second question is how well the representation predicted by our framework can be recovered when only a subset of neurons in the MRNS hidden layer is observed. We find that recovery varies across tasks and depends primarily on the absolute number of recorded neurons, rather than on the sampled fraction of the population alone (Figure S4).

4 Discussion

We developed a theory that derives the geometry and dimensionality of linear networks' minimal representations to solve predictive tasks formulated as transition graphs with *action-labeled* edges. Under our reasonable assumptions, the dominant PCA subspace of such representation comprises task coordinate vectors whose changes depend only on action, allowing its dimension to be computed without network training. Simulations further suggest that cognitive-map geometry reflects not environmental structure alone but its interpretation through task definitions. Edge identities, namely the action labels specifying which transitions are interpreted as instances of the same relation, determine how relationships among experiences are organized, whereas states determine which experiences belong to a common relational system. Our results therefore predict edge identity as a key determinant of studying cognitive-map geometry, particularly in non-spatial tasks, where models based solely on state transitions(Dayan, 1993; Stachenfeld et al., 2017; Mark et al., 2020) usually neglect the relational meaning of edges. Importantly, the resulting geometry is robust to repeated observations and compatible extensions, changing only when state identities or action equivalences are altered.

By translating candidate task models into testable predictions of neural population geometry, the theory also provides a principled way to compare hypotheses about an agent's internal task interpretation. Alternative state and action definitions yield distinct geometries, shifting analysis from decoding experimenter-defined variables to identifying the task organization that best explains neural activity.

Several limitations define the scope of our contributions. First, our theoretical results apply to linear networks with finite, discrete state and action spaces. Although the nonlinear-network simulations show that the predicted representation geometry can remain detectable in learned representations, such observations are not yet supported by an interpretable mathematical theory of nonlinear solutions. Second, the framework assumes deterministic transitions and does not address how transition uncertainty shapes representation geometry. Extending the theory to probabilistic environments remains an important direction for future work. Third, our theory is a normative rather than a generative model of neural representations. The credibility of the broader insights derived from our simulations rests on the network's faithful recovery of the known representation geometry of benchmark tasks. These simulations offer principled hypotheses about cognitive-map representations but their correspondence to neural population activity remains to be tested experimentally.

AI use statement

Generative AI tools were used to assist with grammatical and stylistic polishing of the manuscript and with the development and refinement of analysis code and scientific visualizations. All AI-assisted text, code, analyses, and figures were reviewed and verified by the authors.

Reproducibility statement

The principal results are generated from analytical solutions determined by the explicitly specified state–action transition structures, making the core geometric and dimensionality predictions directly reproducible without neural network training. The appendices provide complete proofs, task definitions, and numerical procedures used in the paper. All code required to reproduce the analyses and figures in this paper is provided in the Supplementary Material.

References

Ege Altan, Sara A Solla, Lee E Miller, and Eric J Perreault. Estimating the dimensionality of the manifold underlying multi-electrode neural recordings. *PLoS computational biology*, 17(11): e1008591, 2021.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

Timothy EJ Behrens, Timothy H Muller, James CR 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.

Lukas Braun, Erin Grant, and Andrew M Saxe. Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. In *International Conference on Machine Learning*, 2025.

György Buzsáki, Sam McKenzie, and Lila Davachi. Neurophysiology of remembering. *Annual review of psychology*, 73:187–215, 2022.

Wilka Carvalho, Momchil S Tomov, William de Cothi, Caswell Barry, and Samuel J Gershman. Predictive representations: Building blocks of intelligence. *Neural Computation*, 36(11):2225–2298, 2024.

Yusi Chen, Huanqiu Zhang, Mia Cameron, and Terrence Sejnowski. Predictive sequence learning in the hippocampal formation. *Neuron*, 112(15):2645–2658, 2024.

Chanwoo Chun, Abdulkadir Canatar, SueYeon Chung, and Daniel Lee. Estimating dimensionality of neural representations from finite samples. In *International Conference on Learning Representations*, volume 2026, pp. 93949–93979, 2026.

Alexandra O Constantinescu, Jill X O'Reilly, and Timothy EJ Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468, 2016.

Hristos S Courellis, Juri Minxha, Araceli R Cardenas, Daniel L Kimmel, Chrystal M Reed, Taufik A Valiante, C Daniel Salzman, Adam N Mamelak, Stefano Fusi, and Ueli Rutishauser. Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 632(8026): 841–849, 2024.

Christopher J. Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *International Conference on Learning Representations*, 2018.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural computation*, 5(4):613–624, 1993.

Lorena Deuker, Jacob LS Bellmund, Tobias Navarro Schröder, and Christian F Doeller. An event map of memory space in the hippocampus. *elife*, 5:e16534, 2016.

Matthew Farrell, Stefano Recanatesi, and Eric Shea-Brown. From lazy to rich to exclusive task representations in neural networks and neural codes. *Current opinion in neurobiology*, 83:102780, 2023.

Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7):1258–1270, 2022.

Spencer Frei, Gal Vardi, Peter L. Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky ReLU networks trained on high-dimensional data. In *International Conference on Learning Representations*, 2023.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *BioRxiv*, pp. 214262, 2017.

Mona M Garvert, Raymond J Dolan, and Timothy EJ Behrens. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *elife*, 6:e17086, 2017.

Dileep George, Rajeev V Rikhye, Nishad Gothoskar, J Swaroop Guntupalli, Antoine Dedieu, and Miguel Lázaro-Gredilla. Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nature communications*, 12(1):2392, 2021.

Kei M Igarashi, Jason Y Lee, and Heechul Jun. Reconciling neuronal representations of schema, abstract task structure, and categorization under cognitive maps in the entorhinal-hippocampal-frontal circuits. *Current opinion in neurobiology*, 77:102641, 2022.

Ari E Kahn, Dani S Bassett, and Nathaniel D Daw. Trial-by-trial learning of successor representations in human behavior. *PLoS computational biology*, 21(11):e1013696, 2025.

Daniel Levenstein, Aleksei Efremov, Roy Henha Eyono, Adrien Peyrache, and Blake Richards. Sequential predictive learning is a unifying theory for hippocampal representation and replay. *BioRxiv*, 2024.

Shirley Mark, Rani Moran, Thomas Parr, Steve W Kennerley, and Timothy EJ Behrens. Transferring structural knowledge across cognitive maps in humans and models. *Nature communications*, 11 (1):4783, 2020.

Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, 2020.

Edward H Nieh, Manuel Schottdorf, Nicolas W Freeman, Ryan J Low, Sam Lewallen, Sue Ann Koay, Lucas Pinto, Jeffrey L Gauthier, Carlos D Brody, and David W Tank. Geometry of abstract learned knowledge in the hippocampus. *Nature*, 595(7865):80–84, 2021.

Yael Niv. Learning task-state representations. *Nature neuroscience*, 22(10):1544–1553, 2019.

Nils Nyberg, Éléonore Duvelle, Caswell Barry, and Hugo J Spiers. Spatial goal coding in the hippocampal formation. *Neuron*, 110(3):394–422, 2022.

Seongmin A Park, Douglas S Miller, Hamed Nili, Charan Ranganath, and Erie D Boorman. Map making: constructing, combining, and inferring on abstract cognitive maps. *Neuron*, 107(6): 1226–1238, 2020.

Brad E Pfeiffer and David J Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74–79, 2013.

Stefano Recanatesi, Gabriel Koch Ocker, Michael A Buice, and Eric Shea-Brown. Dimensionality in recurrent spiking networks: Global trends in activity and local origins in connectivity. *PLoS computational biology*, 15(7):e1006446, 2019.

Stefano Recanatesi, Matthew Farrell, Guillaume Lajoie, Sophie Deneve, Mattia Rigotti, and Eric Shea-Brown. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature communications*, 12(1):1417, 2021.

Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497 (7451):585–590, 2013.

Veronika Samborska, James L Butler, Mark E Walton, Timothy EJ Behrens, and Thomas Akam. Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nature neuroscience*, 25(10):1314–1326, 2022.

Rylan Schaeffer, Mikail Khona, and Ila Fiete. No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *Advances in neural information processing systems*, 35:16052–16067, 2022.

Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. The hippocampus as a predictive map. *Nature neuroscience*, 20(11):1643–1653, 2017.

Christoph Stöckl, Yukun Yang, and Wolfgang Maass. Local prediction-learning in high-dimensional spaces enables neural networks to plan. *Nature Communications*, 15(1):2344, 2024.

Wenbo Tang, Justin D Shin, and Shantanu P Jadhav. Geometric transformation of cognitive maps for generalization across hippocampal-prefrontal circuits. *Cell reports*, 42(3), 2023.

Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

James CR Whittington, David McCaffary, Jacob JW Bakermans, and Timothy EJ Behrens. How to build a cognitive map. *Nature neuroscience*, 25(10):1257–1272, 2022.

Michał J Wójcik, Jake P Stroud, Dante Wasmuht, Makoto Kusunoki, Mikiko Kadohisa, Mark J Buckley, Rui Ponte Costa, Nicholas E Myers, Laurence T Hunt, John Duncan, et al. Learning shapes neural geometry in the primate prefrontal cortex. *Nature Neuroscience*, pp. 1–10, 2026.

Jungsun Yoo, Elizabeth R Chrastil, and Aaron M Bornstein. Cognitive graphs: Representational substrates for planning. *Decision*, 11(4):537, 2024.

Ipshita Zutshi, Athina Apostolelli, Wannan Yang, Zheyang Sam Zheng, Tora Dohi, Edoardo Balzani, Alex H Williams, Cristina Savin, and György Buzsáki. Hippocampal neuronal activity is aligned with action plans. *Nature*, 639(8053):153–161, 2025.

APPENDIX

A PROOFS FOR THE SPECTRAL TRANSFER

A.1 COVARIANCE MATRIX OF AN MRNS REPRESENTATION IN THE SVD BASIS

Let $C_P := I_P - \frac{1}{P}\mathbf{1}_P\mathbf{1}_P^\top \in \mathbb{R}^{P\times P}$ and $H_c := HC_P$. We use the unnormalized covariance matrix $H_cH_c^\top$; the usual factor $\frac{1}{P-1}$ rescales its eigenvalues without changing its eigenvectors. By Theorem 2.2, $H = R\sqrt{N}\,O^\top$ and $R^\top R = I_r$. Therefore,

$$\begin{aligned} H_cH_c^\top &= R\sqrt{N}\,O^\top C_P C_P^\top O\sqrt{N}\,R^\top \\ &= R\sqrt{N}\,O^\top\left(I_P - \frac{1}{P}\mathbf{1}_P\mathbf{1}_P^\top\right)O\sqrt{N}\,R^\top \\ &= R\left[N - \frac{1}{P}\left(\sqrt{N}\,O^\top\mathbf{1}_P\right)\left(\sqrt{N}\,O^\top\mathbf{1}_P\right)^\top\right]R^\top. \end{aligned} \tag{1}$$

Let $a \in \mathbb{R}^r$ be a unit eigenvector of the bracketed matrix with eigenvalue μ:

$$\left[N - \frac{1}{P}\left(\sqrt{N}\,O^\top\mathbf{1}_P\right)\left(\sqrt{N}\,O^\top\mathbf{1}_P\right)^\top\right]a = \mu a, \qquad \|a\|_2 = 1.$$

Since $R^\top R = I_r$,

$$\begin{aligned} H_cH_c^\top(Ra) &= R\left[N - \frac{1}{P}\left(\sqrt{N}\,O^\top\mathbf{1}_P\right)\left(\sqrt{N}\,O^\top\mathbf{1}_P\right)^\top\right](R^\top R)a \\ &= R\left[N - \frac{1}{P}\left(\sqrt{N}\,O^\top\mathbf{1}_P\right)\left(\sqrt{N}\,O^\top\mathbf{1}_P\right)^\top\right]a \\ &= \mu(Ra), \\ \|Ra\|_2^2 &= a^\top R^\top Ra = a^\top a = 1. \end{aligned}$$

Thus, Ra is a unit PCA eigenvector with eigenvalue μ. Define the scalar $\gamma \in \mathbb{R}$ by

$$\gamma := \frac{1}{P}\left(\sqrt{N}\,O^\top\mathbf{1}_P\right)^\top a. \tag{2}$$

This eigenproblem is equivalent to

$$(N - \mu I_r)a = \gamma\sqrt{N}\,O^\top\mathbf{1}_P, \qquad \mu = a^\top Na - P\gamma^2. \tag{3}$$

When $\gamma = 0$, this eigenproblem reduces to the simple form $Na = \mu a$; when $\gamma \neq 0$, the nonhomogeneous term remains.

A.2 THE CASE $\gamma \neq 0$

First suppose $\gamma \neq 0$. If μ is not a singular value of Z, then $a = \gamma(N - \mu I_r)^{-1}\sqrt{N}\,O^\top\mathbf{1}_P \in \mathbb{R}^r$. Substitution into the definition of γ in equation 2 gives the secular equation

$$1 = \frac{1}{P}\sum_{j=1}^{r}\frac{n_j(o_j^\top\mathbf{1}_P)^2}{n_j - \mu}. \tag{4}$$

For each root μ, the corresponding eigenvector is obtained from the expression for a and then normalized.

If instead $\mu = \nu$ is a singular value, let $J_\nu := \{j : n_j = \nu\}$ denote the indices of the diagonal entries of N equal to ν. The component equation $(n_j - \mu)a_j = \gamma\sqrt{n_j}\,o_j^\top\mathbf{1}_P$ in equation 4 requires $o_j^\top\mathbf{1}_P = 0$ for every $j \in J_\nu$. Under this condition, the components outside J_ν are $a_j = \gamma\sqrt{n_j}(o_j^\top\mathbf{1}_P)/(n_j - \nu)$, while the components inside J_ν are free subject to normalization, and consistency requires

$$1 = \frac{1}{P}\sum_{j\notin J_\nu}\frac{n_j(o_j^\top\mathbf{1}_P)^2}{n_j - \nu}. \tag{5}$$

Finally, because $\gamma \neq 0$, equation 3 implies $\mu = a^\top Na - P\gamma^2 < a^\top Na \leq \sigma_{\max}$. Thus every solution in this case satisfies $\mu < \sigma_{\max}$.

A.3 THE CASE $\gamma = 0$

Now suppose $\gamma = 0$. In this case, equation 3 reduces to

$$Na = \mu a, \qquad \left(\sqrt{N}\, O^\top \mathbf{1}_P\right)^\top a = 0.$$

Both equations must hold simultaneously. For a singular value ν of Z, let $E_\nu := \operatorname{span}\{e_j : n_j = \nu\} \subseteq \mathbb{R}^r$, where e_j is the jth standard basis vector of $\mathbb{R}^r$. Since $Na = \nu a$ if and only if $a \in E_\nu$, the solutions with $\mu = \nu$ are exactly the unit vectors in

$$E_\nu \cap \left(\sqrt{N}\, O^\top \mathbf{1}_P\right)^\perp.$$

If ν has multiplicity m_ν, this solution space has dimension m_ν when $o_j^\top \mathbf{1}_P = 0$ for all j with $n_j = \nu$, and dimension $m_\nu - 1$ otherwise. For a simple singular value n_j, a solution exists if and only if $o_j^\top \mathbf{1}_P = 0$.

A.4 PROOFS OF THE SPECTRAL-TRANSFER RESULTS

For a singular value $\sigma > 0$ of Z, define its right singular subspace as $\mathcal{V}_\sigma := \operatorname{span}\{o_j : n_j = \sigma\} \subseteq \mathbb{R}^P$.

Lemma A.1. *For every unit vector $v \in \mathcal{V}_\sigma \cap \mathbf{1}_P^\perp$ and every MRNS, $RO^\top v \subseteq \mathbb{R}^{D_h}$ is a unit PCA eigenvector with eigenvalue σ. Its sample coordinates are*

$$H_c^\top RO^\top v = \sqrt{\sigma}\, v.$$

Theorem A.2. *Suppose that the largest singular value $\sigma_{\max}$ has multiplicity $m > 1$ and that $\mathbf{1}_P \in \mathcal{V}_{\sigma_{\max}}$. Then, for every MRNS, the dominant PCA subspace is*

$$RO^\top\left(\mathcal{V}_{\sigma_{\max}} \cap \mathbf{1}_P^\perp\right).$$

It has eigenvalue $\sigma_{\max}$ and dimension $m - 1$. The corresponding sample-coordinate subspace is $\mathcal{V}_{\sigma_{\max}} \cap \mathbf{1}_P^\perp$.

We now apply the preceding spectral classification to prove these two results.

Proof of Lemma A.1. Take any zero-mean unit right singular vector v of Z associated with σ; equivalently, $v \in \mathcal{V}_\sigma \cap \mathbf{1}_P^\perp$ and $\|v\|_2 = 1$. Express v in the right singular-vector basis by setting $a := O^\top v$. Since $v \in \mathcal{V}_\sigma$, we have $v = Oa$, $\|a\|_2 = 1$, and $Na = \sigma a$. Moreover,

$$\left(\sqrt{N}\, O^\top \mathbf{1}_P\right)^\top a = \sqrt{\sigma}\, \mathbf{1}_P^\top O a = \sqrt{\sigma}\, \mathbf{1}_P^\top v = 0.$$

Thus a is exactly a $\gamma = 0$ solution from Section A.3, with $\mu = \sigma$. The covariance reduction in Section A.1 therefore shows that $Ra = RO^\top v$ is a unit PCA eigenvector with eigenvalue σ. Its sample-coordinate vector is

$$\begin{aligned} H_c^\top Ra &= C_P O \sqrt{N}\, R^\top Ra \\ &= C_P O \sqrt{N}\, a \\ &= \sqrt{\sigma}\, C_P v \\ &= \sqrt{\sigma}\, v. \end{aligned} \tag{6}$$

Thus proves both claims. □

Proof of Theorem A.2. Because $\mathbf{1}_P \in \mathcal{V}_{\sigma_{\max}}$ and $\dim \mathcal{V}_{\sigma_{\max}} = m$, the subspace $\mathcal{V}_{\sigma_{\max}} \cap \mathbf{1}_P^\perp$ has dimension $m - 1$. For $v \in \mathcal{V}_{\sigma_{\max}} \cap \mathbf{1}_P^\perp$, set $a := O^\top v$. Then the orthogonality condition in Section A.3 follows directly:

$$\left(\sqrt{N}\, O^\top \mathbf{1}_P\right)^\top a = \sqrt{\sigma_{\max}}\, \mathbf{1}_P^\top v = 0.$$

By Section A.3 and Lemma A.1, $RO^\top\left(\mathcal{V}_{\sigma_{\max}} \cap \mathbf{1}_P^\perp\right)$ is exactly the dominant PCA eigenspace. □

B PROOFS FOR TASK-SPACE CHARACTERIZATION AND GRAPH DIMENSIONALITY

Lemma B.1. *The set $\mathcal{F}$ of action-consistent task coordinates is a vector subspace of $\mathbb{R}^{D_s}$.*

Proof. The zero coordinate belongs to $\mathcal{F}$ with the zero action-increment vector. Now let $f_1, f_2 \in \mathcal{F}$ have action-increment vectors δ_1, δ_2, respectively. For any $\alpha, \beta \in \mathbb{R}$ and every observed transition (s, a, s'),

$$(\alpha f_1 + \beta f_2)(s') - (\alpha f_1 + \beta f_2)(s) = (\alpha\delta_1 + \beta\delta_2)(a).$$

Thus $\alpha f_1 + \beta f_2$ is action-consistent, with action-increment vector $\alpha\delta_1 + \beta\delta_2$, and hence belongs to $\mathcal{F}$. Therefore $\mathcal{F}$ contains the zero vector and is closed under linear combinations, so it is a vector subspace of $\mathbb{R}^{D_s}$. □

B.1 ACTION-CONSISTENT TASK COORDINATES AND THE MRNS PCA SUBSPACE

Proof of Theorem 2.4. Let $P_X := X^+X \in \mathbb{R}^{P\times P}$, the orthogonal projector onto $\mathcal{R}(X^\top)$, and let f be a unit eigenvector satisfying $ZZ^\top f = cf$. Then $c = f^\top ZZ^\top f$. Since $Z = YP_X$ and P_X is an orthogonal projector, Assumption 1 gives

$$\begin{aligned} f^\top ZZ^\top f &= \|Z^\top f\|_2^2 \qquad (7)\\ &= \|P_X Y^\top f\|_2^2 \\ &= \|Y^\top f\|_2^2 - \|(I_P - P_X)Y^\top f\|_2^2 \\ &= f^\top YY^\top f - \|(I_P - P_X)Y^\top f\|_2^2 \\ &= n_{\text{vis}}\|f\|_2^2 - \|(I_P - P_X)Y^\top f\|_2^2. \qquad (8)\end{aligned}$$

It follows directly that $c \le n_{\text{vis}}$, with equality exactly when

$$c = n_{\text{vis}} \iff (I_P - P_X)Y^\top f = 0 \iff Y^\top f \in \mathcal{R}(X^\top). \qquad (9)$$

The row-space condition is equivalent to

$$Y^\top f \in \mathcal{R}(X^\top) \quad\iff\quad \exists\, u \in \mathbb{R}^{D_s}, t \in \mathbb{R}^{D_a} : \quad Y^\top f = X^\top \begin{bmatrix} u \\ t \end{bmatrix}.$$

For every observed transition (s, a, s'), the one-hot encoding gives $f(s') = u(s) + t(a)$ where $u(s) = e_s^T u$ and $t(a) = e_a^T t$, while Assumption 2 ensures that $f(s) = u(s) + t(a_0)$ at every state. Defining $\delta(a) := t(a) - t(a_0)$ therefore gives $f(s') - f(s) = \delta(a)$ for every observed transition. Conversely, these action-dependent increments imply $Y^\top f = X^\top \left[\begin{smallmatrix} f \\ \delta \end{smallmatrix}\right]$ by the same one-hot encoding. Hence

$$Y^\top f \in \mathcal{R}(X^\top) \quad\iff\quad \exists\, \delta \in \mathbb{R}^{D_a} : \quad Y^\top f = X^\top \begin{bmatrix} f \\ \delta \end{bmatrix}. \qquad (10)$$

Because $Y^\top f \in \mathcal{R}(X^\top)$, then $P_X Y^\top f = Y^\top f$, and therefore

$$ZZ^\top f = YP_X Y^\top f = YY^\top f = n_{\text{vis}} f.$$

Together with equation 9 and equation 10, this proves the null-space characterization in the theorem. A nonzero constant coordinate assignment satisfies the increment condition with $\delta = 0$, so n_{vis} is an eigenvalue of $ZZ^\top$. Since equation 8 gives $c \le n_{\text{vis}}$ for every eigenvalue c, it follows that $\lambda_{\max}(ZZ^\top) = n_{\text{vis}}$ and $\sigma_{\max}(Z) = \sqrt{n_{\text{vis}}}$.

Finally, let $f \in \mathcal{F}$ be a unit vector. By the result above, f is a left singular vector of Z with singular value $\sqrt{n_{\text{vis}}}$. Its corresponding right singular vector is

$$o_f = \frac{1}{\sqrt{n_{\text{vis}}}} Z^\top f = \frac{1}{\sqrt{n_{\text{vis}}}} Y^\top f.$$

o_f is normalized because $\|Y^\top f\|_2^2 = f^\top Y Y^\top f = n_{\text{vis}}$. Assumption 1 also implies $Y\mathbf{1}_P = n_{\text{vis}}\mathbf{1}_{D_s}$, and hence

$$o_f^\top \mathbf{1}_P = \frac{1}{\sqrt{n_{\text{vis}}}} f^\top Y \mathbf{1}_P = \sqrt{n_{\text{vis}}}\, f^\top \mathbf{1}_{D_s},$$

Taking f to be the normalized constant coordinate assignment shows that $\mathbf{1}_P \in \mathcal{V}_{\sigma_{\max}}$. Now let $f \in \mathcal{F} \cap \mathbf{1}_{D_s}^\perp$ have unit norm. The identities above give $o_f = Y^\top f/\sqrt{n_{\text{vis}}}$ and $o_f^\top \mathbf{1}_P = 0$. Together with Theorem A.2, this identifies these directions with the dominant PCA subspace of every MRNS. Setting $v = o_f$ by Lemma A.1 therefore gives the sample-coordinate vector along the direction $RO^\top o_f$:

$$H_c^\top RO^\top o_f = \sqrt{\sigma_{\max}}\, o_f = n_{\text{vis}}^{-1/4} Y^\top f,$$

where we used $\sigma_{\max} = \sqrt{n_{\text{vis}}}$. Hence, for an observed transition (s_i, a_i, s_i'), its coordinate along $RO^\top o_f$ is

$$\big(H_c^\top RO^\top o_f\big)_i = n_{\text{vis}}^{-1/4} f(s_i') = n_{\text{vis}}^{-1/4}\big(f(s_i) + \delta_f(a_i)\big).$$

Thus, up to the common scale $n_{\text{vis}}^{-1/4}$, $f(s_i')$ is exactly the sample's coordinate along $RO^\top o_f$, and $\delta_f(a_i)$ is the displacement along that coordinate produced by action a_i. A different orthonormal basis within the repeated maximal singular subspace merely rotates these coordinate directions and leaves the retained sample geometry unchanged. □

B.2 Graph-nullity formula

Theorem B.2. *Under Assumptions 1– 3,*

$$d_{\text{task}} = \dim \mathcal{N}\big(\begin{bmatrix} Y^\top - S^\top & -A^\top \end{bmatrix}\big) - 1.$$

Proof of Theorem B.2. By the block form of X, the matrix identity in equation 10 is equivalent to

$$\begin{bmatrix} Y^\top - S^\top & -A^\top \end{bmatrix} \begin{bmatrix} f \\ \delta \end{bmatrix} = 0. \tag{11}$$

By the definition of $\mathcal{F}$, projection of $\mathcal{N}\big(\begin{bmatrix} Y^\top - S^\top & -A^\top \end{bmatrix}\big)$ onto its first D_s coordinates has image $\mathcal{F}$. Under Assumption 3, $AA^\top$ has full rank, so $\mathcal{N}(A^\top) = \{0\}$.

Suppose that $\left[\begin{smallmatrix} f \\ \delta_1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} f \\ \delta_2 \end{smallmatrix}\right]$ both belong to the null space in equation 11. Their constraint equations are

$$(Y^\top - S^\top) f = A^\top \delta_1, \qquad (Y^\top - S^\top) f = A^\top \delta_2. \tag{12}$$

Subtracting gives $A^\top(\delta_1 - \delta_2) = 0$, so $\mathcal{N}(A^\top) = \{0\}$ implies $\delta_1 = \delta_2$. Therefore, each $f \in \mathcal{F}$ uniquely determines its action-increment vector δ.

Consequently,

$$\dim \mathcal{F} = \dim \mathcal{N}\big(\begin{bmatrix} Y^\top - S^\top & -A^\top \end{bmatrix}\big). \tag{13}$$

By definition, $d_{\text{task}} = \dim \mathcal{F} - 1$. Substituting equation 13 gives the stated graph-nullity formula. Because the constraint matrix is constructed entirely from the observed transitions, the result is independent of the choice of MRNS. □

B.3 Connectivity–increment decomposition and inverse-action bound

Proof of Theorem 2.5. By Assumption 3, $AA^\top$ is invertible and every $f \in \mathcal{F}$ has a unique increment vector δ_f, as shown in the proof of Theorem B.2. Multiplying $(Y^\top - S^\top) f = A^\top \delta_f$ by A gives

$$A(Y^\top - S^\top) f = AA^\top \delta_f, \qquad \delta_f = (AA^\top)^{-1} A(Y^\top - S^\top) f.$$

Define

$$D := (AA^\top)^{-1} A(Y^\top - S^\top).$$

As f ranges over $\mathcal{F}$, the vectors $Df = \delta_f$ are exactly the globally compatible increment vectors. Therefore

$$d_{\text{inc}} = \dim \mathcal{R}(D|_{\mathcal{F}}).$$

The rank–nullity theorem applied to $D|_{\mathcal{F}}$ gives

$$\dim \mathcal{F} = \dim \mathcal{N}(D|_{\mathcal{F}}) + \dim \mathcal{R}(D|_{\mathcal{F}}).$$

We next determine $\mathcal{N}(D|_{\mathcal{F}})$. For an observed transition (s, a, s'), the corresponding entry of $(Y^\top - S^\top)f$ is

$$\big((Y^\top - S^\top)f\big)_i = f(s'_i) - f(s_i).$$

Therefore $(Y^\top - S^\top)f = 0$ exactly when f is constant on each connected component. If their vertex sets are $V_1, \dots, V_G$, then the corresponding indicator functions are

$$\mathbf{1}_{V_j}(s) := \begin{cases} 1, & s \in V_j, \\ 0, & s \notin V_j, \end{cases} \qquad j = 1, \dots, G.$$

They have disjoint supports and form a basis of the component-wise constant functions. Hence

$$\mathcal{N}(Y^\top - S^\top) = \operatorname{span}\{\mathbf{1}_{V_1}, \dots, \mathbf{1}_{V_G}\}, \qquad \dim \mathcal{N}(Y^\top - S^\top) = G.$$

For every $f \in \mathcal{N}(Y^\top - S^\top)$, choosing $\delta = 0$ satisfies $(Y^\top - S^\top)f = A^\top \delta$, so $\mathcal{N}(Y^\top - S^\top) \subseteq \mathcal{F}$. Therefore,

$$\mathcal{N}\big((Y^\top - S^\top)|_{\mathcal{F}}\big) = \mathcal{F} \cap \mathcal{N}(Y^\top - S^\top) = \mathcal{N}(Y^\top - S^\top).$$

For $f \in \mathcal{F}$, the vector $(Y^\top - S^\top)f$ lies in $\mathcal{R}(A^\top)$. The invertibility of $AA^\top$ implies that A is injective on this range. Together with the definition of D, this gives

$$Df = 0 \quad \Longleftrightarrow \quad A(Y^\top - S^\top)f = 0 \quad \Longleftrightarrow \quad (Y^\top - S^\top)f = 0.$$

It follows that $\mathcal{N}(D|_{\mathcal{F}}) = \mathcal{N}((Y^\top - S^\top)|_{\mathcal{F}}) = \mathcal{N}(Y^\top - S^\top)$, so $\dim \mathcal{N}(D|_{\mathcal{F}}) = G$.

Together with $d_{\text{inc}} = \dim \mathcal{R}(D|_{\mathcal{F}})$, rank–nullity now yields

$$\dim \mathcal{F} = G + d_{\text{inc}}, \qquad d_{\text{task}} = \dim \mathcal{F} - 1 = G - 1 + d_{\text{inc}},$$

Finally, index the p inverse-action pairs by (a_k, a_k^{-1}). For every $f \in \mathcal{F}$, a self-transition labelled by the stay action gives zero increment, while inverse closure gives the opposite increment along the reverse transition:

$$\delta_f(a_0) = 0, \qquad \delta_f(a_k^{-1}) = -\delta_f(a_k), \quad k = 1, \dots, p.$$

Therefore every vector in $\mathcal{R}(D|_{\mathcal{F}})$ has the form

$$[0, \alpha_1, -\alpha_1, \dots, \alpha_p, -\alpha_p]^\top.$$

These vectors lie in a p-dimensional subspace of $\mathbb{R}^{2p+1}$, so

$$d_{\text{inc}} = \dim \mathcal{R}(D|_{\mathcal{F}}) \le p, \qquad d_{\text{task}} \le G + p - 1.$$

□

B.4 Examples: Effects of path relations and action compatibility on task dimensionality

To illustrate how path relations can tighten this bound, suppose the observed transitions include

$$\text{Father} \xrightarrow{a_{FM}} \text{Mother}, \qquad \text{Mother} \xrightarrow{a_{MS}} \text{Son}, \qquad \text{Father} \xrightarrow{a_{FS}} \text{Son},$$

together with their inverse transitions. For every $f \in \mathcal{F}$, the direct and two-step paths from Father to Son must have the same net increment. Hence

$$\delta_f(a_{FS}) = f(\text{Son}) - f(\text{Father}) = \delta_f(a_{FM}) + \delta_f(a_{MS}).$$

Thus the three positive action coordinates satisfy one linear relation and contribute at most two independent increment directions. In this example $p = 3$ but $d_{\text{inc}} \le 2 < p$, showing explicitly how path relations can make the action contribution strictly smaller than its nominal upper bound.

Action compatibility can reduce task dimensionality even when the underlying state space is unchanged. Consider an otherwise standard connected 2D grid task, whose nonconstant compatible task coordinates are spanned by the two spatial coordinates. Write such a task coordinate as $f(x, y) = ux + vy$. Suppose that the same action label a is incorrectly assigned to three transitions with displacements $(2, 1)$, $(1, 2)$ and $(1, 3)$. Action consistency then requires

$$\delta_f(a) = 2u + v = u + 2v = u + 3v,$$

and hence $u = v = 0$. The two coordinate directions therefore vanish, reducing d_{task} from 2 to 0. Thus an action definition that groups incompatible state changes can remove otherwise valid task dimensions.

C TASK DEFINITION

In every task, states and actions are categorical identities. For an observed transition (s, a, s'), the network input is $[e_s^{(D_s)}; e_a^{(D_a)}]$ and the target is $e_{s'}^{(D_s)}$. Coordinates described below define the transition graph but are not included in the input. Repeated transitions used for target balancing retain the same state and action codes.

2D/3D navigation task. In the 2D task, each state is one location in a 4×5 grid. The actions are stay, left, right, up, and down. A movement transition is included only when the adjacent target location lies inside the grid. In the 3D task, each state is one occupied voxel in four stacked square layers of sizes 7×7, 5×5, 3×3, and 1×1, giving 84 states. The actions are stay and movement by one voxel in either direction along the x, y, or z axis; a movement transition is included only when its target voxel exists.

Family-relation task. Each state is one of 13 people in a four-generation pedigree. The pedigree contains the direct line of descent, collateral branches through sons and daughters, and spouses who enter from outside the bloodline. Every recorded spouse or parent–child relation is represented in both directions. The action labels are stay, wife, husband, father-to-son, son-to-father, father-to-daughter, daughter-to-father, mother-to-son, son-to-mother, mother-to-daughter, and daughter-to-mother. Thus, people in the same relational role share an action label even when they belong to different generations or branches.

Disconnected-map task. The 34 states comprise two disconnected, topologically identical maze maps with 17 locations each. Each map consists of three parallel five-location arms joined at one end by a five-location transverse corridor, with shared junctions counted only once. Map identity is part of state identity, and no transition joins the two maps. Each map permits stay, left, right, up, and down transitions wherever the corresponding adjacent location exists. In the shared-action task, one common set of five action labels is used across both maps. In the map-specific task, each map has its own set of five labels, so, for example, "right in map 1" and "right in map 2" are different actions.

Garvert relational graph task. The 12 states are the 12 object identities at the vertices of the graph from Garvert et al. (2017). In the edge-specific task, every directed graph edge $i \to j$ has its own action a_{ij}, together with stay. In the merged-action task, directed edges with the same one of six orientations in the triangular layout share an action label; stay is unchanged. These labels are analytical encodings and were not actions presented in the original experiment.

City–family superposition task. The states are the 20 locations of a 4×5 grid. The 13 people are assigned to 13 fixed, distinct locations. Navigation actions connect adjacent grid locations. A family relation from person p to person p' becomes a transition from the location assigned to p to the location assigned to p', with the same family-action label as above. Navigation and family relations therefore act on the same location identities.

City–family Cartesian product task. Each state is a pair consisting of one of the 20 grid locations and one of the 13 people, giving 260 states. A navigation action changes the location and leaves the person unchanged. A family action changes the person and leaves the location unchanged. The navigation and family action vocabularies are combined, with one shared stay label.

The nonlinear-network and neuron-subsampling analyses reuse these same (X, Y) task encodings and introduce no additional state or action definitions.

D SIMULATION METHODS

D.1 ACTION-MERGING NETWORK

We tested whether the initially distinct actions in the edge-specific Garvert task naturally merge under an information bottleneck, and whether the emergent action organization corresponds to the six

direction classes defined by the graph's minimum-dimensional Euclidean triangular layout. These layout-derived classes were used only for evaluation and were never provided to the network.

For each hidden width $D_h \in \{1, \dots, 10\}$, we trained a bias-free two-layer linear network on the 84 target-balanced transition samples,

$$H = W_1 X, \qquad \widehat{Y} = W_2 H, \qquad \mathcal{L}_{\text{pred}} = \frac{1}{84} \|\widehat{Y} - Y\|_F^2, \tag{14}$$

where X concatenated 12-dimensional state and 49-dimensional action one-hot codes (48 edge-specific actions and one shared stay action), and Y encoded the next state. We used Xavier-uniform initialization with scale 0.01, because small initialization more closely approaches the theoretical solution, and optimized the prediction loss by full-batch SGD with learning rate 0.01 for 50,000 epochs. Each width was repeated for 100 random seeds.

Writing $W_1 = [U\ V]$ for its state and action blocks gives $h(s, a) = u_s + v_a$. We defined each edge action's displacement relative to stay as

$$\Delta h_a = h(s, a) - h(s, \text{stay}) = v_a - v_{\text{stay}}, \tag{15}$$

and measured the mean cosine similarity between action pairs assigned to the same layout-defined direction, opposite directions, or distinct non-opposite directions. Figure S1 reports the mean and one standard deviation across seeds.

D.2 Successor representation

For the 12-state Garvert graph, we defined the row-stochastic transition matrix P by $P_{ij} = \Pr(S_{t+1} = j \mid S_t = i)$ and computed the successor representation

$$M_\gamma = (I - \gamma P)^{-1} = \sum_{k=0}^{\infty} \gamma^k P^k, \tag{16}$$

whose rows index current states and whose columns index discounted future-state occupancies. Figure S2 used $\gamma \in \{0.1, 0.2, \dots, 0.9, 0.99\}$ and compared three constructions. In the random-walk condition (RW), the agent moved uniformly among the neighbours of each state, with no stay transition. In the diagonal-compensation control (DC), we added $\text{diag}(r)$ to the RW successor matrix after its computation, where, in state order $0, \dots, 11$, $r = (3, 3, 2, 0, 2, 3, 0, 0, 3, 3, 2, 3)$; this control is not itself the successor representation of a Markov chain. In the balanced-visitation condition (BV), each outgoing edge, one base stay, and the same additional stay counts r_i were treated as transition-frequency observations. Thus every state had seven samples, with $P_{ii} = (1 + r_i)/7$ and $P_{ij} = 1/7$ for each neighbour j.

Each construction was analysed using either neuron-wise centered PCA or neuron-wise z-scored PCA, yielding the six conditions RW-centered, RW-z-scored, DC-centered, DC-z-scored, BV-centered, and BV-z-scored. States were observations and successor neurons were features; z-scoring used the sample standard deviation across the 12 states after column centering. For display only, each PC1–PC2 embedding was orthogonally Procrustes-aligned to the fixed triangular layout; this rotation or reflection did not change distances, explained variance, or the PCA subspace.

D.3 Nonlinear network

For each task's input and output (X, Y), we trained a bias-free network with $H_{\text{pre}} = W_1 X$, $H_{\text{post}} = \phi(H_{\text{pre}})$, and $\widehat{Y} = W_2 H_{\text{post}}$, where ϕ was ReLU or sigmoid. The loss was

$$\mathcal{L}_z = \frac{1}{P} \left[\|W_2 \phi(W_1 X) - Y\|_F^2 + \lambda \left(\|H_z\|_F^2 + \|W_2\|_F^2 \right) \right], \qquad z \in \{\text{pre}, \text{post}\}. \tag{17}$$

We used $\lambda = 0$ without regularization and $\lambda = 0.01$ for either choice of z. Duplicate transitions were multiplicity weighted. Hidden width was $N_h = 2\,\text{rank}(Z)$: 36 (2D grid), 168 (3D grid), 26 (family), 68 (each disconnected-grid task), 512 (factorized composition), and 24 (each Garvert task). Initialization scales were $\{0.01, 0.1, 1, 5\}$; Xavier uniform was used for sigmoid networks and W_2, and Kaiming uniform for ReLU W_1. We used 64-bit Adam (learning rate 0.01), at most $100{,}000$ steps, and five seeds unless noted otherwise.

PCA was performed on H_{pre}. Let $A, B \in \mathbb{R}^{P\times d}$ be the unwhitened analytical-MRNS and learned-network scores for the same transition samples, where $d = d_{\mathrm{task}}$, and center their columns to obtain A_c and B_c. We used

$$\mathrm{sim}(A, B) := \max_{Q^\top Q = I_d} \frac{\langle A_c, B_c Q\rangle_F}{\|A_c\|_F \|B_c\|_F} = \frac{\|B_c^\top A_c\|_*}{\|A_c\|_F \|B_c\|_F}, \tag{18}$$

where $\|\cdot\|_*$ is the nuclear norm. If $B_c^\top A_c = U\Sigma V^\top$, then $Q^* = UV^\top$ and the numerator is $\mathrm{tr}(\Sigma)$. The score is scale invariant, lies in $[0, 1]$, and permits rotations and reflections. Scores were neither whitened nor averaged over samples sharing a next state.

All analytical and learned-network PC score matrices were computed once for all transition samples and held fixed during permutation testing; only their sample correspondence was randomized. For Table S1, similarity was computed using the first d_{task} learned-network PCs. For Table S2, the highest similarity over all d_{task}-PC subsets among the first 20 PCs was used; the search covered all 17 PCs for the rank-17 Garvert merged-action networks.

Within each seed, we generated 1,000 independent global sample-label permutations. The leading-PC set remained fixed, whereas the best-PC search was repeated for every permutation. Seed-level similarities were averaged by condition using independent permutations across seeds, and one-sided p values used the plus-one correction. Benjamini–Hochberg correction was applied to the 192 conditions separately for each table, whose symbols report the resulting q values.

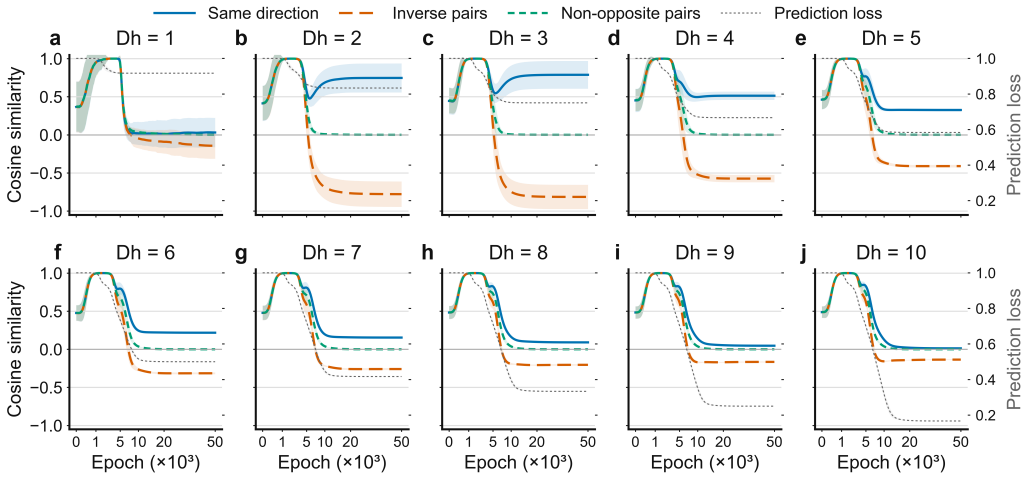


Figure S1: **Action merging weakens as the information bottleneck relaxes.** Training dynamics for linear networks with hidden width $D_h = 1$–10. Curves show means; colored bands show one standard deviation across 100 random seeds.

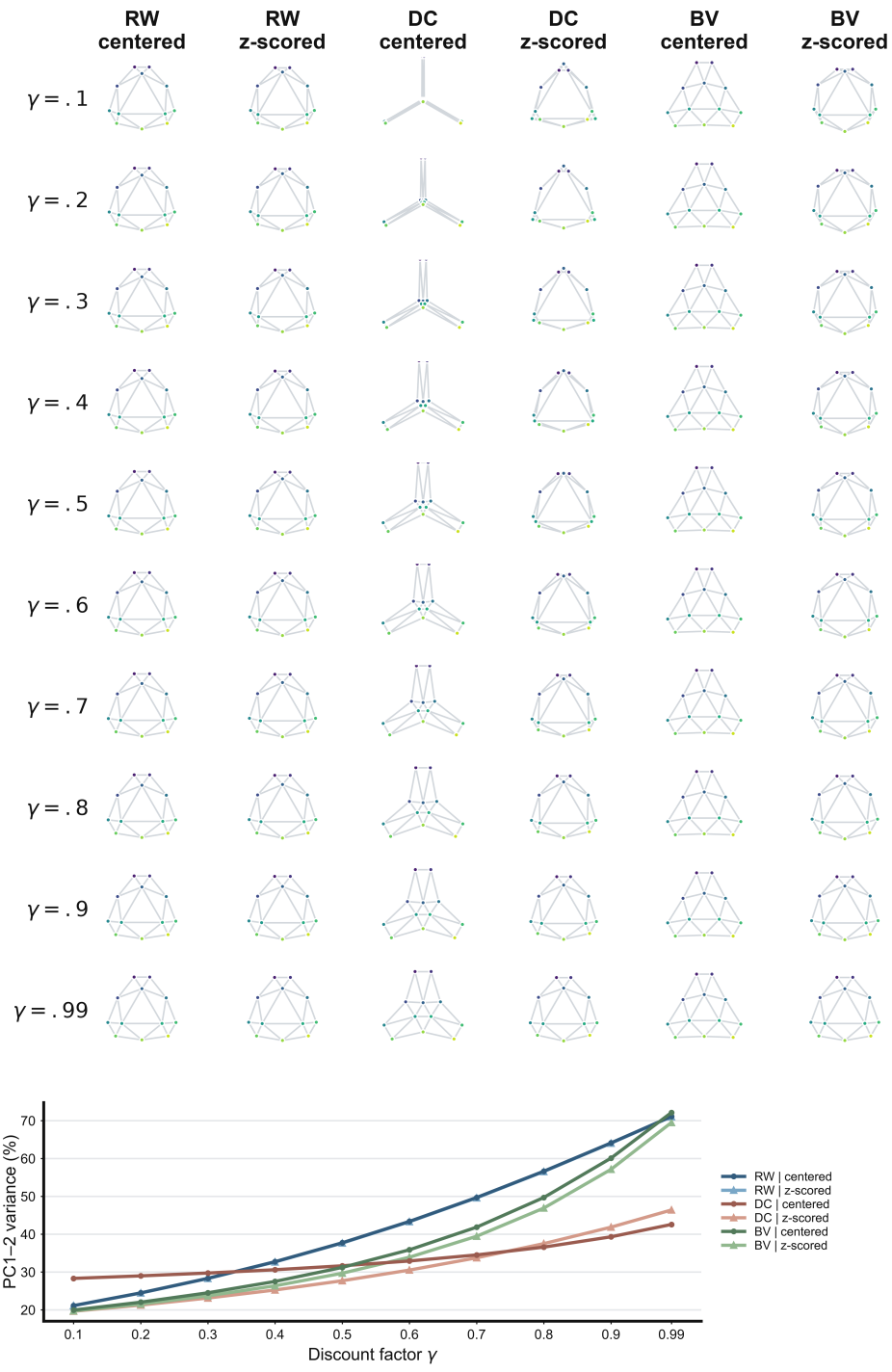


Figure S2: **Successor-representation geometry depends on policy, scaling, and discount factor.** PCA embeddings of the Garvert-graph successor representation across discount factors $\gamma = 0.1$–0.99. Columns compare neuron-wise centered and neuron-wise z-scored PCA under a uniform random-walk policy (RW), a post-SR diagonal-compensation control (DC), and the balanced-visitation policy (BV) used in the main text; rows vary γ. Gray lines show graph adjacency and colors identify states. The lower panel shows the variance jointly explained by PC1 and PC2.

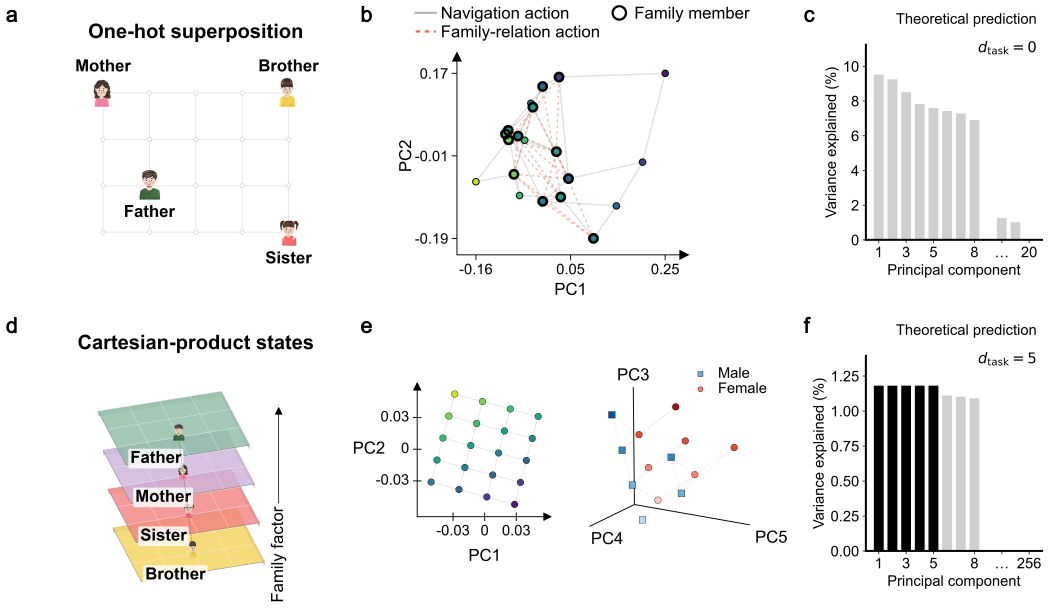


Figure S3: **Factorized states enable distinct relational systems to coexist within a single representation. a–c**, Schematic, MRNS representation, and explained-variance spectrum of the one-hot superposition task, respectively. **d–f**, Corresponding results for the Cartesian-product state task; for visualization, the five-dimensional subspace in **e** is orthogonally rotated into two city and three family axes.

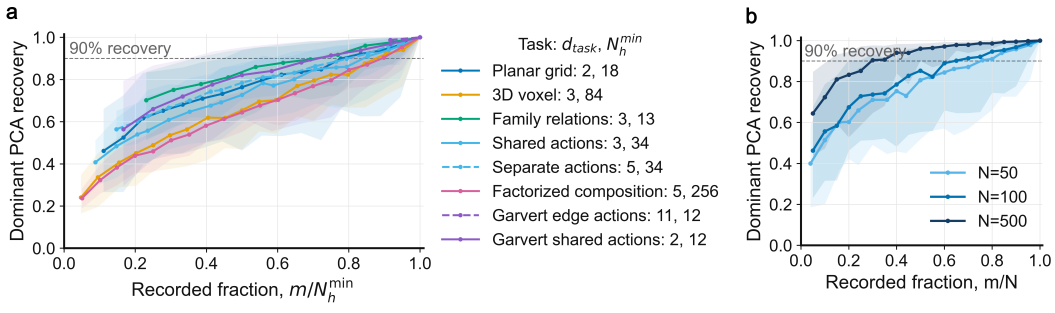


Figure S4: **Dominant MRNS coordinates are recoverable from partial neural recordings. a**, Normalized orthogonal-Procrustes similarity between the leading unwhitened PCA coordinates from each complete minimum-width MRNS population and from m sampled neurons. For each sampling level, neural mixing and neuron subsampling were independently repeated 80 times. Before PCA, the covariance estimated from the m sampled neurons was multiplied by N_h/m to compensate for the expected reduction in overall scale. Curves show mean recovery, and shading shows the empirical 2.5th–97.5th percentiles across repeats. **b**, Corresponding recovery for the two-dimensional grid embedded in populations of $N = 50$, 100, and 500 neurons.

Table S1: **Leading-PC recovery across nonlinear-network training conditions.**

	Init. 0.01		Init. 0.1		Init. 1		Init. 5	
Task	ReLU	Sigmoid	ReLU	Sigmoid	ReLU	Sigmoid	ReLU	Sigmoid
No regularization								
2D grid	0.606^{**}	0.451^{**}	0.450^{**}	0.334^{**}	0.287^{**}	0.308^{**}	0.248^{**}	0.238^{**}
3D grid	0.106^{**}	0.097^{**}	0.078^{**}	0.097^{**}	0.085^{**}	0.072^{*}	0.135^{**}	0.161^{**}
Family relations	0.729^{**}	0.664^{**}	0.646^{**}	0.521^{**}	0.353^{**}	0.455^{**}	0.260^{**}	0.239^{**}
Disconnected maps, shared actions	0.389^{**}	0.363^{**}	0.304^{**}	0.328^{**}	0.287^{**}	0.223^{**}	0.220^{**}	0.230^{**}
Disconnected maps, map-specific actions	0.757^{**}	0.536^{**}	0.606^{**}	0.476^{**}	0.481^{**}	0.335^{**}	0.424^{**}	0.344^{**}
Factorized composition	0.399^{**}	0.179^{**}	0.377^{**}	0.205^{**}	0.322^{**}	0.247^{**}	0.149^{**}	0.171^{**}
Garvert graph, merged actions	0.597^{**}	0.220^{**}	0.356^{**}	0.290^{**}	0.221^{**}	0.328^{**}	0.228^{**}	0.254^{**}
Garvert graph, edge-specific actions	0.822^{**}	0.989^{**}	0.651^{**}	0.990^{**}	0.500^{**}	0.975^{**}	0.397^{**}	0.401^{**}
Pre-activation regularization of H_{pre} and W_2								
2D grid	0.699^{**}	0.371^{**}	0.578^{**}	0.365^{**}	0.416^{**}	0.378^{**}	0.489^{**}	0.398^{**}
3D grid	0.128^{**}	$\mathbf{1.000}^{**}$	0.120^{**}	$\mathbf{1.000}^{**}$	0.183^{**}	$\mathbf{1.000}^{**}$	0.153^{**}	$\mathbf{0.995}^{**}$
Family relations	0.863^{**}	0.657^{**}	0.836^{**}	0.551^{**}	0.647^{**}	0.598^{**}	0.711^{**}	0.551^{**}
Disconnected maps, shared actions	0.438^{**}	0.970^{**}	0.423^{**}	0.980^{**}	0.408^{**}	0.976^{**}	0.413^{**}	0.328^{**}
Disconnected maps, map-specific actions	0.416^{**}	0.496^{**}	0.442^{**}	0.558^{**}	0.351^{**}	0.552^{**}	0.329^{**}	0.213^{**}
Factorized composition	0.406^{**}	0.899^{**}	0.343^{**}	$\mathbf{1.000}^{**}$	0.356^{**}	$\mathbf{0.997}^{**}$	0.194^{**}	$\mathbf{0.541}^{**}$
Garvert graph, merged actions	0.788^{**}	0.616^{**}	0.773^{**}	0.584^{**}	0.624^{**}	0.641^{**}	0.633^{**}	0.206^{**}
Garvert graph, edge-specific actions	1.000^{**}	1.000^{**}	1.000^{**}	1.000^{**}	1.000^{**}	1.000^{**}	1.000^{**}	1.000^{**}
Post-activation regularization of H_{post} and W_2								
2D grid	0.679^{**}	0.717^{**}	0.522^{**}	0.662^{**}	0.418^{**}	0.527^{**}	0.276^{**}	0.351^{**}
3D grid	0.096^{**}	0.161^{**}	0.093^{**}	0.153^{**}	0.101^{**}	0.105^{**}	0.085^{**}	0.094^{**}
Family relations	0.901^{**}	0.936^{**}	0.820^{**}	0.926^{**}	0.499^{**}	0.912^{**}	0.296^{**}	0.898^{**}
Disconnected maps, shared actions	0.425^{**}	0.410^{**}	0.378^{**}	0.420^{**}	0.352^{**}	0.383^{**}	0.219^{**}	0.397^{**}
Disconnected maps, map-specific actions	0.576^{**}	0.417^{**}	0.487^{**}	0.432^{**}	0.410^{**}	0.410^{**}	0.393^{**}	0.421^{**}
Factorized composition	0.406^{**}	0.523^{**}	0.385^{**}	0.521^{**}	0.321^{**}	0.276^{**}	0.161^{**}	0.102^{**}
Garvert graph, merged actions	0.880^{**}	0.926^{**}	0.797^{**}	0.853^{**}	0.422^{**}	0.772^{**}	0.178^{*}	0.647^{**}
Garvert graph, edge-specific actions	0.939^{**}	0.956^{**}	0.921^{**}	0.959^{**}	0.636^{**}	0.958^{**}	0.394^{**}	0.959^{**}

Table S2: **Best-PC recovery across nonlinear-network training conditions.**

	Init. 0.01		Init. 0.1		Init. 1		Init. 5	
Task	ReLU	Sigmoid	ReLU	Sigmoid	ReLU	Sigmoid	ReLU	Sigmoid
No regularization								
2D grid	0.726^{**}	0.580^{**}	0.627^{**}	0.586^{**}	0.566^{**}	0.583^{**}	0.488^{**}	0.522^{**}
3D grid	0.417^{**}	0.388^{**}	0.417^{**}	0.389^{**}	0.382^{**}	0.370^{**}	0.256^{**}	0.282^{**}
Family relations	0.810^{**}	0.736^{**}	0.765^{**}	0.664^{**}	0.568^{**}	0.613^{**}	0.542^{**}	0.546^{**}
Disconnected maps, shared actions	0.693^{**}	0.561^{**}	0.659^{**}	0.571^{**}	0.606^{**}	0.537^{**}	0.450^{**}	0.470^{**}
Disconnected maps, map-specific actions	0.787^{**}	0.702^{**}	0.716^{**}	0.657^{**}	0.611^{**}	0.574^{**}	0.561^{**}	0.474^{**}
Factorized composition	0.507^{**}	0.562^{**}	0.507^{**}	0.560^{**}	0.459^{**}	0.420^{**}	0.277^{**}	0.295^{**}
Garvert graph, merged actions	0.807^{**}	0.683^{**}	0.747^{**}	0.680^{**}	0.601^{**}	0.699^{**}	0.548^{**}	0.596^{**}
Garvert graph, edge-specific actions	0.857^{**}	$\underline{0.989}^{**}$	0.777^{**}	$\underline{0.990}^{**}$	0.634^{**}	$\underline{0.975}^{**}$	0.455^{**}	0.636^{**}
Pre-activation regularization of H_{pre} and W_2								
2D grid	$\underline{0.699}^{**}$	0.549^{**}	0.644^{**}	0.513^{**}	0.616^{**}	0.560^{**}	0.607^{**}	0.521^{**}
3D grid	0.438^{**}	$\underline{\mathbf{1.000}}^{**}$	0.460^{**}	$\underline{\mathbf{1.000}}^{**}$	0.489^{**}	$\underline{\mathbf{1.000}}^{**}$	0.405^{**}	$\underline{\mathbf{0.995}}^{**}$
Family relations	0.918^{**}	0.790^{**}	0.895^{**}	0.792^{**}	0.792^{**}	0.764^{**}	0.810^{**}	0.725^{**}
Disconnected maps, shared actions	0.635^{**}	$\underline{0.970}^{**}$	0.619^{**}	$\underline{0.980}^{**}$	0.619^{**}	$\underline{0.976}^{**}$	0.618^{**}	0.489^{**}
Disconnected maps, map-specific actions	0.656^{**}	0.786^{**}	0.669^{**}	0.804^{**}	0.580^{**}	0.821^{**}	0.583^{**}	0.641^{**}
Factorized composition	0.588^{**}	$\underline{0.899}^{**}$	0.566^{**}	$\underline{\mathbf{1.000}}^{**}$	0.554^{**}	$\underline{\mathbf{0.997}}^{**}$	0.395^{**}	$\mathbf{0.621}^{**}$
Garvert graph, merged actions	0.886^{**}	0.813^{**}	0.930^{**}	0.747^{**}	0.896^{**}	0.798^{**}	0.890^{**}	0.643^{**}
Garvert graph, edge-specific actions	$\underline{1.000}^{**}$	$\underline{1.000}^{**}$	$\underline{1.000}^{**}$	$\underline{1.000}^{**}$	$\underline{1.000}^{**}$	$\underline{1.000}^{**}$	$\underline{1.000}^{**}$	$\underline{1.000}^{**}$
Post-activation regularization of H_{post} and W_2								
2D grid	0.723^{**}	$\underline{0.717}^{**}$	0.628^{**}	0.706^{**}	0.566^{**}	0.661^{**}	0.472^{**}	0.512^{**}
3D grid	0.457^{**}	0.660^{**}	0.429^{**}	0.635^{**}	0.377^{**}	0.652^{**}	0.248^{**}	0.610^{**}
Family relations	$\underline{0.901}^{**}$	$\underline{0.936}^{**}$	0.893^{**}	$\underline{0.926}^{**}$	0.665^{**}	$\underline{0.912}^{**}$	0.518^{**}	0.913^{**}
Disconnected maps, shared actions	0.671^{**}	0.654^{**}	0.632^{**}	0.650^{**}	0.544^{**}	0.611^{**}	0.446^{**}	0.597^{**}
Disconnected maps, map-specific actions	0.733^{**}	0.689^{**}	0.708^{**}	0.663^{**}	0.608^{**}	0.667^{**}	0.523^{**}	0.643^{**}
Factorized composition	0.518^{**}	0.625^{**}	0.512^{**}	0.642^{**}	0.459^{**}	0.647^{**}	0.284^{**}	0.612^{**}
Garvert graph, merged actions	0.902^{**}	$\underline{0.926}^{**}$	0.871^{**}	0.861^{**}	0.642^{**}	0.859^{**}	0.583^{**}	0.821^{**}
Garvert graph, edge-specific actions	$\underline{0.939}^{**}$	$\underline{0.956}^{**}$	$\underline{0.921}^{**}$	$\underline{0.959}^{**}$	0.714^{**}	$\underline{0.958}^{**}$	0.450^{**}	$\underline{0.959}^{**}$

Entries are mean normalized orthogonal-Procrustes similarities across seeds and unaveraged transition samples. Table S1 uses PCs $1, \ldots, d_{\mathrm{task}}$; Table S2 exhaustively selects the best d_{task}-PC subset among the first $m = \min(20, r)$ nonzero PCs ($m = 17$ for Garvert merged actions; otherwise $m = 20$). Underlined S2 entries indicate that the optimal subset was PCs $1, \ldots, d_{\mathrm{task}}$ for every seed, matching the S1 subset. Significance used 1,000 label permutations per seed, repeating S2 selection per permutation, and Benjamini–Hochberg correction across 192 conditions per table ($^{*}q < 0.05$, $^{**}q < 0.01$). Bold marks conditions in which every network's prediction MSE differed from the nonzero least-squares baseline $L_{\mathrm{LS}} = \|YX^{+}X - Y\|_F^2/P$ by at most $10^{-3}L_{\mathrm{LS}}$. Unless noted, $n = 5$ seeds.