# Cross-individual translation of spontaneous zebrafish brain activity through a shared latent representation

Mattéo Dommanget-Kott[1,2], Jorge Fernandez-de-Cossio-Diaz[3], Guillaume Faye-Bédrin[1,4], Georges Debrégeas[1+*], Volker Bormuth[1+*]

**1** Institut de Biologie Paris-Seine, Laboratoire Jean Perrin, Sorbonne Université, CNRS, Paris, France
**2** Université Paris Cité, Paris, France
**3** Institut de Physique Théorique, Université Paris-Saclay, CNRS, CEA, Gif-sur-Yvette, France
**4** Laboratoire de Physique de l'École Normale Supérieure, CNRS, PSL, Sorbonne Université, Université de Paris, Paris, France

+ Senior authors with equal contribution
* georges.debregeas@sorbonne-universite.fr
* volker.bormuth@sorbonne-universite.fr

## Abstract

Spontaneous activity is a hallmark of brain function, reflecting the underlying circuit organization. Identifying conserved structure across individuals in this self-sustained activity has remained a longstanding challenge, especially in vertebrates where one-to-one neuron correspondence is inaccessible. Here, we introduce latent-aligned Restricted Boltzmann Machines (LaRBMs), an unsupervised generative approach that uncovers a common representational space from cell-resolved whole-brain recordings in larval zebrafish. This latent space consists of spatially localized co-activation motifs, or cell assemblies, that generalize across animals and form interpretable building blocks of population-wide activity. LaRBMs enable bidirectional mapping of brain states between individuals: activity patterns from one fish can be encoded into the latent space and decoded into another. The translated patterns are assigned high probability by the recipient model and retain the original spatial organization. These results show that spontaneous activity in the vertebrate brain is highly stereotyped at the level of functional cell assemblies and can be reliably captured through a common latent code. Because it provides an interpretable and quantitative framework for functional cross-individual alignment, LaRBM paves the way for comparative phenotyping of brain activity across developmental, genetic, and pathological variation.

## Significance Statement

Spontaneous brain activity, without external stimuli, shapes development, constrains coding, and reflects neural organization. Whether this activity reveals similar organization across individuals has been unclear. Using single-cell, whole-brain recordings in larval zebrafish and statistical learning, we find that spontaneous dynamics share a common structure across animals. Spatially organized co-activated neuron assemblies recur across individuals and act as building blocks of population activity. This shared representation enables fictive translation of activity from one fish

into another's neural space while preserving spatial and statistical plausibility. These results suggest that brains organize population activity by similar principles to represent internal states. Our findings reveal conserved organization in the vertebrate brain and establish a quantitative framework for comparing brain dynamics across individuals and conditions.

## Key words

Spontaneous brain activity; Cross-individual alignment; Population codes; Latent variable models; Restricted Boltzmann Machine; Generative models; Neural assemblies; Shared latent space; Zebrafish whole-brain imaging.

## Introduction

Spontaneous brain activity is a fundamental feature of neural systems, shaping sensory processing, behavior, and internal state dynamics even in the absence of external stimuli [10, 23, 4]. In humans, population-level analyses of resting-state fMRI have revealed large-scale networks that are spatially and functionally conserved across individuals [9]. These so-called resting-state networks (RSNs) have been interpreted as a manifestation of intrinsic brain organization and are increasingly used to study brain development [40], aging [14], and psychiatric disorders [5]. However, their interpretability is limited by low spatial and temporal resolution.

At the other end of the spectrum, exhaustive recordings in small organisms, such as *C. elegans*, have enabled dynamic models that directly link activity to connectome structure. These models can predict transitions between behavioral states and reveal low-dimensional manifolds underlying brain-wide activity [3, 13, 2]. However, they are largely restricted to small nervous systems with stereotyped anatomy and limited neuron count.

In between these two extremes lies a major challenge: how to model spontaneous activity at single-cell resolution in vertebrates with non-stereotyped, complex architectures spanning multiple scales, while retaining interpretability and enabling cross-individual comparisons. Zebrafish larvae offer a unique opportunity to fill this gap. Their optical transparency and compact brains enable functional imaging of nearly all neurons at single-cell resolution over extended periods [31, 1]. Recent studies have leveraged this animal model to study sensory-motor transformations [43, 26, 30], behavioral state dynamics [24, 29], spontaneous brain activity [41, 35, 33], or internal models [32].

Approaches to inter-individual comparison in neuroscience have mainly relied on anatomical registration combined with stimulus- or behavior-guided voxel-wise averaging or clustering. Although useful, these methods assume voxel-level correspondence and overlook the rich variabilities at neuronal scales. More recent frameworks such as Shared Response Models (SRMs) [7], hyperalignment [16, 15, 20, 21], spatial autoencoders [6], and contrastive frameworks for cross-subject alignment [36] aim to learn shared embeddings across subjects. However, most analyses remain limited to individual subjects or rely on anatomical alignment combined with stimulus- or behavior-guided averaging, making them unsuitable for extracting shared structure from spontaneous activity at the neuronal level.

Probabilistic generative energy-based models offer a powerful alternative [17]. They have been used to model retinal activity, hippocampal replay [8, 34], and patterns of spontaneous activity [41, 27]. In particular, Restricted Boltzmann Machines (RBM) model the joint distribution of neural activity via latent variables (hidden units) that

can be interpreted as co-activating neuronal assemblies. However, these studies have generally focused on single datasets, and it remains unclear whether RBMs can be used to identify conserved structures across individuals, especially when applied to full-brain recordings at cellular resolution.

Here, we introduce a framework for discovering a shared, interpretable latent structure in spontaneous brain-wide activity across individuals. Using whole-brain calcium imaging in larval zebrafish, we train RBMs with a shared latent space, a strategy we term Latent-aligned RBMs (LaRBMs), in which the hidden units are constrained to maintain their prior functional organization and activation distribution across individuals. This enables us to extract a common set of latent motifs, corresponding to spatially organized assemblies, that are reused across different brains. Crucially, we go beyond statistical comparison by projecting activity from one fish into the shared latent space and decoding it into the neural space of another fish. We show that spontaneous patterns at single cell resolution can be mapped across individuals, while preserving both spatial organization and statistical structure.

Our approach fills a key gap between connectome-informed modeling in C. elegans and population-level RSN analysis in humans. It enables comparative, interpretable, and probabilistically grounded modeling of spontaneous brain-wide activity across individuals. This provides a concrete route for identifying shared circuit motifs that shape spontaneous dynamics at the mesoscale, bringing us closer to a common representational framework for a vertebrate brain.

# Results

## Latent representations of spontaneous brain-wide activity are variable and non-alignable across individual RBMs

We showed previously in van der Plas et al. [41] that spontaneous whole-brain neuronal activity of zebrafish larvae could be modeled by Restricted Boltzmann Machines (RBMs). In this framework, the binarized spiking activity of individual neurons is mapped onto a latent space comprised of hidden units that correspond to cell assemblies (*i.e.* functional ensembles of co-activating neurons; Fig. 1 A). Each hidden unit $\mu$ is characterized by the spatial distribution of weights $w_{i\mu}$, where $i$ is the neuronal index. These weights are found to be sparse, spatially compact, and define functional circuits and anatomical structures. This biological interpretability suggests that RBMs trained on recordings from different fish would converge on similar representations, which could then be used to perform cross-individual comparisons. However, as RBM training is stochastic by nature, we expect a certain variability in this representation.

**RBMs represent neuronal statistics from a degenerate set of latent features** To assess the variability in representation we investigated the robustness of the latent features by training multiple RBMs on the same neuronal dataset (fish 1 in Tab. 1). We used the same training parameters and selected the models that had reached a high level of statistical performance (see Sec. Teacher RBMs and Sec. Evaluation).

We then computed the Pearson correlation $\rho_{\mu\nu} = \rho(w_{i\mu}, w'_{i\nu})$ between the weight matrices $w_{i\mu}$ and $w'_{i\nu}$ of two trainings. For the best alignment, we found that $\approx 28\%$ of hidden units could be unequivocally paired between trainings ($\rho_{\mu\nu} > 0.9$), while $\approx 34\%$ had no matching pair ($\rho_{\mu\nu} < 0.2$) (see Fig. 1 B-C). This indicates that only a core set of hidden units are stable across trainings.

We performed the same analysis on the expected hidden activity $h_\mu(t) = \mathbb{E}[h_\mu \mid \mathbf{v}(t)]$ by computing the correlation $\rho_{\mu\nu} = \rho(h_\mu(t), h_\nu(t))$, and found that $\approx 54\%$ of hidden units had correlations $\rho_{\mu\nu} > 0.9$ and $\approx 18\%$ $\rho_{\mu\nu} < 0.2$ (see Fig. S1). This result reflects

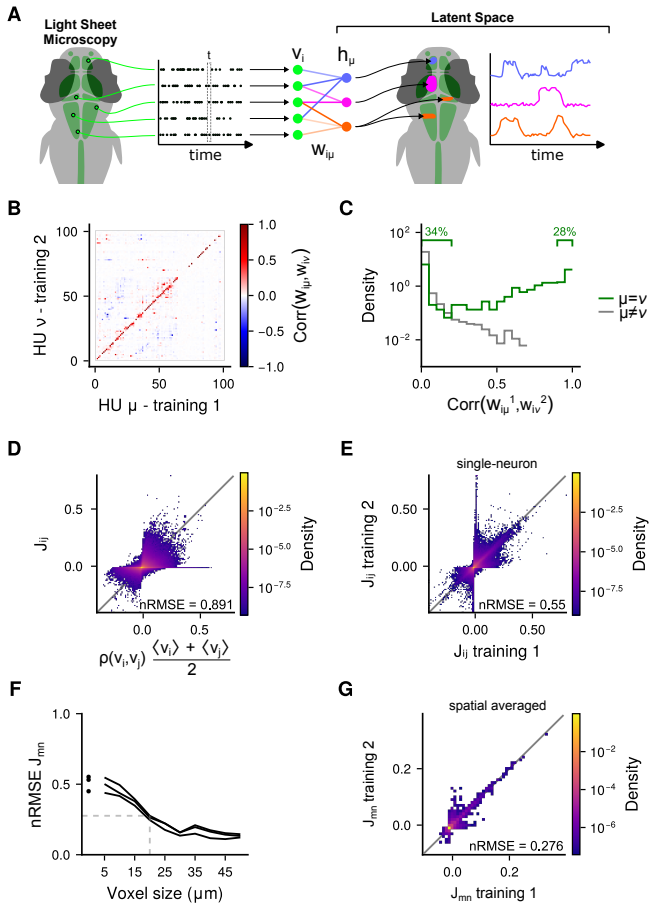



**Fig 1. Restricted Boltzmann Machines (RBMs) produce degenerate representations of neuronal functional organization. A**: Zebrafish larvae neuronal activity is recorded using Light Sheet Fluorescence Microscopy and deconvolved into binarized spike trains. RBMs are trained with binary visible units $v_i$ corresponding to neurons and connected to dReLU hidden units $h_\mu$ via a weight matrix $w_{i\mu}$ (see Methods Restricted Boltzmann Machines). The hidden layer describes a latent space capturing the activation modes of the whole neuronal population as groups of co-activating neurons [41]. **B**: Pairwise Pearson correlation between the weight matrices $w_{i\mu}$ and $w'_{i\nu}$ of two example RBMs trained on the same neuronal recording. The matrix rows have been reordered to obtain the best alignment between the two trainings. **C**: Distribution of correlations from panel B, for best alignment pairs ($\nu = \mu$, green) and other pairs ($\nu \neq \mu$, grey). **D**: Rescaled pairwise Pearson correlation $\rho(v_i, v_j)\frac{\langle v_i\rangle+\langle v_j\rangle}{2}$ between neurons $i$ and $j$ vs. coupling matrix $J_{ij}$ (see Eq.14) inferred from an example RBM. **E**: Comparison of two example coupling matrices inferred from two RBMs trained on the same neuronal recording. **F**: nRMSE (see Methods Measuring identity) between coupling matrices inferred from three different training on the same neuronal recording. Dots correspond to $J_{ij}$ matrices as presented in panel C. Lines correspond to coupling matrices spatially averaged on a cubic-voxel grid : $J_{mn} = \frac{1}{|\mathcal{V}_m|\ |\mathcal{V}_n|}\sum_{i\in\mathcal{V}_m}\sum_{j\in\mathcal{V}_n} J_{ij}$ with $\mathcal{V}_m$ and $\mathcal{V}_n$ the sets of neurons in voxels $m$ and $n$ respectively. **G**: Comparison of two example coupling matrices $J_{mn}$ spatially-averaged at $20\mu m$ voxel size.

the fact that, despite being connected to different neuronal populations, many hidden units share similar temporal signals, capturing dominant dynamics of brain activity spanning large brain regions.

**RBMs produce stable functional connectivity model only at coarse-grained scales** We showed previously [41] that effective couplings $J_{ij}$ between neurons $i$ and $j$ could be inferred from the model (see Sec. Inferred neuronal couplings), and that this measure of functional connectivity reflects, at least in part, the structural connectivity measured between brain regions. Importantly, $J_{ij}$ is not a measurement of the correlation between neurons, but reflects a model of connectivity which takes into account indirect network interactions (see Fig. 1 D).

We compared the effective couplings $J_{ij}$ inferred from multiple trainings on the same dataset as described above. We found that, while they were strongly correlated, couplings were also highly variable, with a significant fraction of neuron pairs being strongly coupled in one model and completely decoupled in another (Fig. 1 E). However, when couplings were spatially averaged over cubic-voxels of side $\gtrapprox 20\mu m$, the resulting $J_{ij}$ was much more stereotypical between trainings (Fig. 1 F-G). This suggests that, while the inferred couplings between neurons are degenerate between RBM trainings, they are well defined at a coarse-grained scale.

In summary, we found that RBMs provide partially degenerate representations of neuronal activity, reflecting the stochastic nature of the model and its training algorithm, as well as insufficient constraints from the data. These findings caution against naively comparing RBMs trained on different individuals, particularly at neuronal scales. However, the stability of the model at mesoscopic scales suggests that RBMs could capture a stereotypical organization of spontaneous brain activity which might be comparable across individuals, motivating multi-individual training paradigms that explicitly align latent spaces.

## A global voxel-level RBM uncovers shared structure in spontaneous brain activity across individuals

We first asked whether spontaneous whole-brain activity across individual zebrafish larvae could be described by a common latent representation, at a coarse-grain level, using anatomical cross-individual registration.

To enable cross-individual comparisons, we spatially registered six fish to a common brain coordinate system using affine and non-linear transformations (see Methods: Morphological Registration). The resulting aligned spontaneous brain activity from all six fish is visualized side by side in Movie S1, with each fish shown separately to highlight individual dynamics within the shared coordinate space. Each neuron was then assigned to a voxel in a 3D grid (20 $\mu m$ cube size), and voxel activity was computed as the mean of the normalized fluorescence ($\Delta F/F$) of all neurons within it (Fig. 2 A and Methods Voxelisation). Only voxels that contained at least two neurons in each fish were retained, ensuring that anatomical coverage was consistent across individuals. This filtering step yielded a total of 2,995 shared voxels, with an average of $11 \pm 5$ neurons per voxel (Fig. S2 D mean and standard deviation). Importantly, this voxel size was the smallest possible to retain $> 75\%$ of all neurons in all fish, thus balancing spatial resolution with inter-individual consistency (Fig. S2 B).

We then trained an RBM on the concatenated voxelized activity from all fish (Fig. 2 A-B). This global RBM consisted of 2995 Gaussian visible units (corresponding to the number of shared voxels) and 40 hidden units (see Methods Training). Before training, we $z$-scored each voxel's activity within each fish to account for differences in

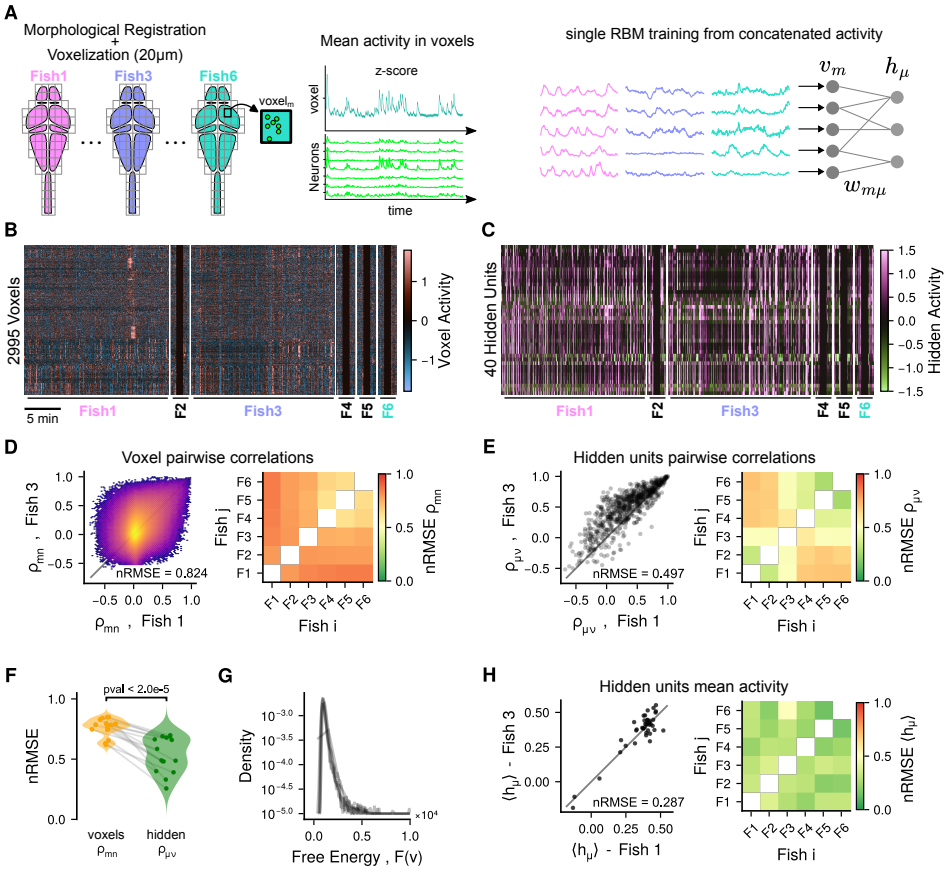



**Fig 2. Concatenation and voxelization allow for a common RBM representation.** **A**: Pipeline of voxelized RBM training. Averaged brain scans are first morphologically registered to a common reference brain. A grid of cubic voxels with side $20\mu m$ is then used to group neurons spatially. The neuronal activity within each voxel is averaged and z-scored. The voxel activity for all 6 fish is concatenated to construct a single dataset, which is then used to train an RBM (see Methods Voxelised RBMs) **B**: Concatenated voxelized activity of all 6 fish. Fish 1 and 3 are shown in full, while the activity of fish 2, 4, 5 and 6 are truncated for visualization purposes. **C**: The activity of panel B *translated* into the hidden space of the RBM (see Methods Mapping from one fish to another). **D**: Stereotypy of the pairwise Person correlation coefficient $\rho_{mn}$ between voxels $m$ and $n$. Left: example fish pair. Right: nRMSE (see Methods Measuring identity) of $\rho_{mn}$ between all pairs of fish. **E**: Stereotypy of the pairwise Person correlation coefficient $\rho_{\mu\nu}$ between hidden units $\mu$ and $\nu$. Left: example fish pair. Right: nRMSE of $\rho_{\mu\nu}$ between all pairs of fish. **F**: Distributions of fish-to-fish nRMSE for pairwise Pearson correlations between voxels (left orange, same as panel D right) and between hidden units (right green, same as panel E right). Each dot represents a pair of fish. The p-value was computed with a one-tailed Mann-Whitney U test. **G**: Distribution of free energy $F(\mathbf{v})$ (see Methods Restricted Boltzmann Machines) of voxel activity configurations $\mathbf{v}$. One line per fish. **H**: Stereotypy of the mean activity $\langle h_\mu \rangle$ of hidden units $\mu$. Left: example fish pair. Right: nRMSE of $\langle h_\mu \rangle$ between all pairs of fish.

mean and variance across individuals (see Methods Voxelisation) [19]. Without this normalization, voxelized training resulted in the RBM capturing fish identity rather

than shared features.

**The global RBM is a good model for data from individual fish.** Validation of this global RBM model on each fish separately demonstrated that the model accurately captured pairwise voxel correlations in 5 out of 6 fish, and more generally, provided a good statistical model of voxelized activity across individuals (see Fig. S4). This indicates that, despite voxel-level variability, the RBM can extract a common statistical structure. This is particularly notable because no fish-specific information was provided during training.

**Latent representations are more stereotyped than raw activity.** Projecting each fish's activity into the RBM's hidden space (Fig. 2 B-C) revealed that correlations between hidden units were more conserved across individuals than correlations between voxels (Fig. 2 D-F). This was quantified using the normalized Root Mean Squared Error (nRMSE). Pairwise voxel correlations had high fish-to-fish variability (nRMSE up to $\sim$0.8) with weak voxel size dependence (Fig. S2 G). In contrast, hidden-unit correlations were significantly more stereotyped (nRMSE $\sim$0.3–0.7, Mann–Whitney U test $p<0.01$). Interestingly, we observed that fish from the same clutch (F1–F3 vs. F4–F6) clustered together in the latent representations (see Fig. 2 G), although no explicit labeling was provided. This suggests that the RBM captures subtle commonalities in spontaneous activity linked to biological or experimental context.

**RBM captures shared structure without encoding individual idiosyncrasies.** To verify that the model did not simply encode for fish-specific features, we considered evaluating the probability $P(\mathbf{v})$ of neuronal configurations. Evaluating $P(\mathbf{v})$ is computationally impractical, and we used the free energy as an alternative:

$$F(\mathbf{v}) = -\log Z - \log P(\mathbf{v}) \tag{1}$$

where $Z$ is the partition function (see Methods Definition). We found that free energy distributions for all fish overlapped closely (Fig. 2 G), indicating that each individual's activity was equally well-explained by the model.

We further compared the mean hidden activity $\langle h_\mu \rangle$ across fish, and found it to be comparable between individuals (Fig. 2 H), suggesting that the RBM did not merely classify individuals.

Overall, these findings confirmed that the latent space was shared and not specialized to any specific fish. The dynamics of activity in this space might be different for different fish, but the overall distribution is similar. As such, it describes the common realm of possible hidden configurations $\mathbf{h}$.

Despite anatomical variability, spontaneous brain activity across zebrafish larvae can be represented in a shared latent space. A single RBM trained on voxelized and concatenated data successfully captures second-order activity statistics across individuals. Importantly, latent representations show greater inter-fish stereotypy than voxelized activity, revealing higher-order regularities in collective dynamics. Crucially, this increased alignment is not trivially explained by dimensionality reduction (see Fig. S3 D) but rather reflects the ability of the RBM to extract conserved co-activation motifs that generalize across animals. These findings establish the feasibility of constructing unified, interpretable representations of spontaneous neural activity across individuals.

## Single-cell resolved LaRBMs reveal conserved spatial cell assemblies via a shared latent space

While voxel-based registration enabled us to construct a common latent space at a coarse spatial resolution, this approach sacrifices single-neuron interpretability. Next, we asked whether we could preserve interpretability while enforcing a shared representation of spontaneous whole-brain activity across individual fish.

We developed a training framework in which a reference RBM, the *teacher*, defines the latent space, and subsequent *student* RBMs are trained on different fish under the constraints that they use the same latent space (Fig. 3 A). Each RBM models the binarized spiking activity of $43480 \pm 2755$ neurons (mean and std, see Materials List of fish) recorded in individual fish at single-cell resolution.

In practice, we first trained the teacher RBM on one fish, with binary visible units and 100 hidden units to define the latent space (Methods Teacher RBMs). Each hidden unit $\mu$ in this trained teacher model represents a distributed functional assembly, a spatial pattern of coactive neurons, as defined by its weight vector $w_{i\mu}$ (see examples in Fig. 3 B and Fig. S7 B). Student RBMs have the same number of hidden units, but can have a different number of visible units to account for variability in neuron-count per fish. To initialize the student RBM parameters, we interpolated the spatial weight patterns from the trained teacher onto the neuron positions in the student brains, we copied the hidden unit potentials $\mathcal{U}_\mu(h)$ of the teacher, and we estimated the fields of the visible units from the corresponding neuron's baseline activity in the student (see Methods Student LaRBM Initialization). These pre-initialized student RBMs do not capture the statistical structure of their respective fish, as they have not yet been trained on the new data. If trained without additional constraints, the student representations would adapt solely to their local data and gradually drift away from the teacher's latent structure. We therefore introduced a new paradigm, which we call latent-aligned RBMs (LaRBMs), in which the student RBM is trained not only to maximize the likelihood of its own observed neuronal activity statistics, but also to capture the distribution of hidden-unit activations sampled from the teacher RBM:

$$\mathcal{L} = (1-\lambda)\,\langle \log P(\mathbf{v}^S_{\text{data}})\,\rangle + \lambda\,\langle \log P(\mathbf{h}^T_{\text{sampled}})\,\rangle \tag{2}$$

where $\mathbf{v}^S_{\text{data}}$ is the neuronal activity of the student, $\mathbf{h}^T_{\text{sampled}}$ are hidden configurations sampled from the teacher RBM, and $\lambda \in [0,1]$ is an hyperparameter controlling the relative impact of each term on training (see Methods Latent-aligned Restricted Boltzmann machines (LaRBMs) for details and Fig. 3 A). This approach ensures that all student RBMs share a subspace of hidden configurations aligned with the teacher, while retaining sufficient flexibility to adapt to individual neuronal datasets.

**LaRBM students converge faster and more reliably.** The LaRBM students were found to converge rapidly, requiring only one-tenth the number of training steps compared to unconstrained and randomly initialized models. This suggests that the initialization step acts as pretraining, while the actual training merely fine-tunes the model to match student statistics. Indeed, we found that, contrary to teachers where multiple RBMs needed to be trained to obtain a good model, student LaRBMs were consistently able to reproduce the first- and second-order statistics of neuronal activity, sometimes even exceeding that of the teachers (see Fig. S5 A-B and Fig. S6). Despite being trained with a latent-space constraint, student RBMs remain faithful to the distribution of neuron firing rates and pairwise covariances in their respective datasets. Notably, even though training does not explicitly account for correlations between hidden units, we found that their pairwise covariances were partially preserved between the teacher and student (see Fig. S5 C-D). This consistency across animals supports the

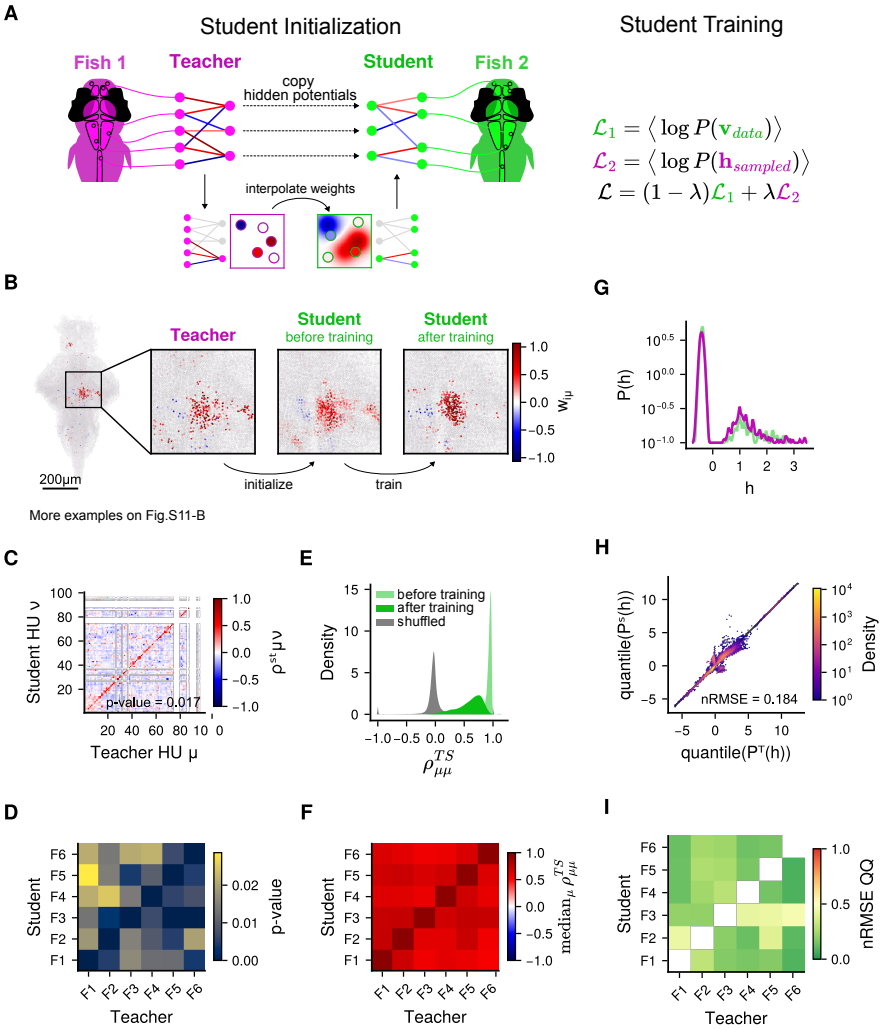



**Fig 3. RBMs can be trained from priors and constrained to enforce a similar hidden space. A**: Teacher/Student paradigm. A *student* LaRBM is initialized from a pretrained *teacher* RBM by copying its hidden layer and spatial interpolation of its weights. The student RBM is then trained on both student neuronal data and hidden configurations sampled from the teacher RBM. **B**: Example spatial interpolation of weights. Left: z-projection map and inset of teacher weights $w^T_{i\mu}$ for each neuron $i$ and a single hidden unit $\mu$. Middle: interpolated weights in student. Right: weights after student training. **C**: Example spatial correlation matrix $\rho^{TS}_{\mu\nu}$ (see Methods Measuring Spatial Similarity) between weights maps of hidden unit $\mu$ of teacher RBM and hidden unit $\nu$ of student RBM. **D**: P-value of diagonal dominance of the $\rho^{TS}_{\mu\nu}$ matrix (see Methods Measuring Diagonal Dominance) for every pair of teacher and student. **E**: Distribution of spatial correlations $\rho^{TS}_{\mu\mu}$ between corresponding hidden units in all teacher-student pairs for students before training (light green), students after training (green), and shuffled pairs of hidden units (grey). **F**: Median of spatial correlation $\text{median}_\mu\, \rho^{TS}_{\mu\mu}$ for every teacher-student pair. **G**: Prior distribution of hidden value $P(h)$ for the example hidden unit presented in panel B, in both teacher (magenta) and trained student (green) RBMs. **H**: Quantile-quantile (Q-Q) plot of $P(h)$ in the example teacher and student pair for all hidden units combined. **I**: nRMSE (see Methods Measuring identity) of Q-Q plot in panel H for every teacher-student pair.

hypothesis that the RBM's hidden units reflect stereotyped functional connectivity motifs.

**Hidden units retain their spatially localized patterns.** Each hidden unit in the teacher RBM represents a spatially localized pattern of co-activating neurons. After interpolating the teacher weights into a student RBM and training the model, we observed that these spatial patterns were largely retained (examples in Fig. 3 B and Fig. S7 B). The spatial correlation between corresponding hidden units in the teacher and student RBMs was on average above 0.6 and significantly higher than between non-matching pairs (see Fig. 3 C-E, bootstrap p-value $\approx 0.02$, see Methods Measuring Spatial Similarity and Measuring Diagonal Dominance), indicating that most hidden units can be reliably matched between models based on their spatial weight patterns. We observed that student weight maps were generally of lower amplitude and more compact, that is, involving fewer strongly connected neurons $N_\mu^S = |w_{i\mu} > 10^{-5}|_i$ with $N_\mu^S \approx \frac{1}{2} N_\mu^T$ (see Fig. S7 A-B). This is consistent with the new likelihood at Eq. 2 and the weight regularization applied during training (see Supplementary Section Small weight expansion of the LaRBM). Crucially, the core spatial identity of the assemblies was maintained.

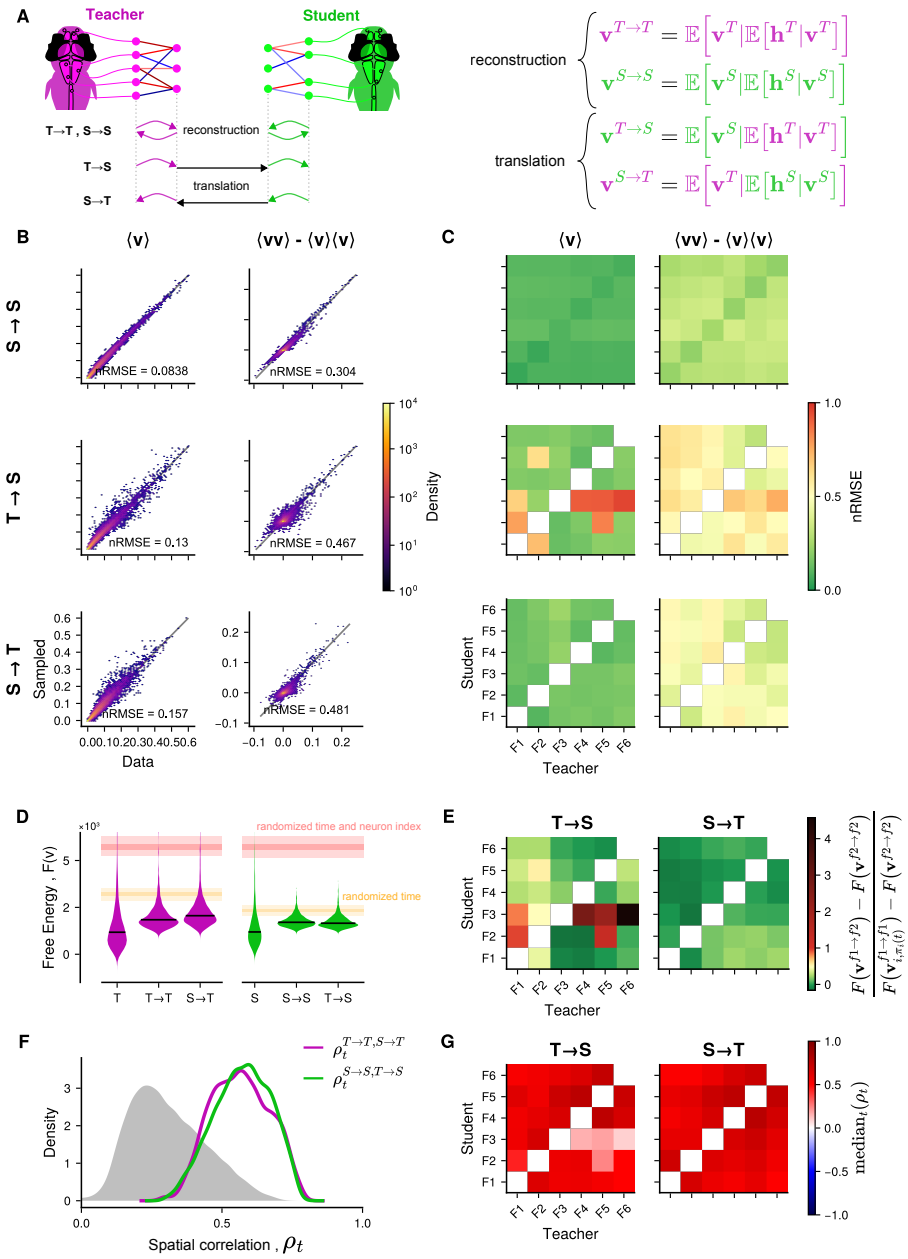
**Student RBMs preserve the teacher's hidden-unit priors.** Finally, we examined whether student RBMs accessed the same regions of the latent space as the teacher. We compared the prior distributions $P(h_\mu)$ of each hidden unit and found close agreement between the teacher and student RBMs (Fig. 3 G-I). Quantile–quantile (Q–Q) plots showed strong linear correspondence (Fig. 3 H), and the identity error (nRMSE) across all teacher-student pairs was low (Fig. 3 I). This indicates that, despite anatomical and statistical variability, spontaneous activity across individuals can be mapped to the same latent sub-space.

Together, these results demonstrate that RBMs trained on different individuals can share a common latent space in which functional assemblies are reliably matched based on spatial structure and prior activation. However, this alone does not guarantee that individual brain states (*i.e.* specific whole-brain spontaneous activity patterns) are encoded similarly across fish.

## Cross-individual translation confirms conserved encoding of spontaneous activity in the shared latent space

Restricted Boltzmann Machines enable encoding of neuronal configurations into the latent space, and decoding of latent configurations back into the neuronal space. This bidirectional generative capacity, combined with our shared latent representations, enables the translation of any activity pattern from one individual to another through the latent space. Here, we evaluate whether such synthetic (translated) activity patterns share the same characteristics as the endogenous brain activity.

Specifically, starting from a neuronal configuration $\mathbf{v}^T$ from a teacher fish, we successively computed the expected hidden activity $\mathbb{E}[h^T \,|\, \mathbf{v}^T]$ and the decoded activity pattern $\mathbb{E}[\mathbf{v}^S \,|\, h^T]$ in the student brain (see Methods Mapping from one fish to another). We then assessed whether this translated activity is highly probable under the student model and retains the same spatial structure as the original teacher pattern. Because the mapping is bidirectional, it allows us to benchmark both teacher-to-student ($T \to S$) and student-to-teacher ($S \to T$) translations against within-fish reconstructions ($T \to T$ or $S \to S$) and randomized controls (Fig. 4 A). We show in Movie S3 and Movie S4 two example movies of $\mathbf{v}^{T\to S}$ and $\mathbf{v}^{S\to T}$, respectively, to illustrate this activity translation.




*(Caption on next page.)*

**Fig 4. Spontaneous neuronal activity can be translated through the latent space to another fish. A**: Neuronal configurations can be represented in hidden space and then either reconstructed in the original neuronal space, or translated to another neuronal space. (see Methods Mapping from one fish to another) **B**: Neuronal statistics for the example teacher-student pair. Mean neuronal activity $\langle v \rangle$ (left) and pairwise covariance $\langle v_i v_j \rangle - \langle v_i \rangle \langle v_j \rangle$ (right) between empirical data and reconstructed student configurations ($S \to S$, top) and translated configurations in both teacher-to-student ($T \to S$, middle) and student-to-teacher ($S \to T$, bottom). **C**: nRMSE (see Methods Measuring identity) for each panel of B and every teacher-student pair. **D**: Distributions of free energy for empirical ($T$ or $S$), reconstructed ($T \to S$ or $S \to T$), and translated ($S \to T$ or $T \to S$) configurations in an example teacher (left) and student (right) RBMs. Black lines indicate median. Horizontal bands indicate 50% and 99% of free energy distribution for randomized data with shuffled time frames (orange), and shuffled time frames and neurons (red). **E**: Median free energy contrast $\frac{F(\mathbf{v}^{f_1 \to f_2}) - F(\mathbf{v}^{f_2 \to f_2})}{F(\mathbf{v}^{f_1 \to f_1}_{i,\pi_i(t)}) - F(\mathbf{v}^{f_2 \to f_2})}$ of translated configurations ($\mathbf{v}^{f_1 \to f_2}$) with respect to reconstructed configurations ($\mathbf{v}^{f_2 \to f_2}$) and time-randomized configurations ($\mathbf{v}^{f_1 \to f_1}_{i,\pi_i(t)}$). **F**: Spatial correlation $\rho_t^{T \to T, S \to T}$ (magenta) and $\rho_t^{S \to S, T \to S}$ (green) between reconstructed and translated neuronal configuration for each time frame of an example teacher-student pair. In gray, the distribution of spatial correlations between shuffled pairs of time frames. **G**: Median of the distributions of spatial correlations from panel F for every teacher-student pair.

Qualitatively, these movies show that this procedure is indeed capable of generating activity which resembles real brain activity, with translated patterns recapitulating salient features of spontaneous brain-wide activity, including spatially organized assemblies and bilateral coordination.

**Statistical features are preserved during mapping.** We first compared the mean activity $\langle v \rangle$ and pairwise covariances $\langle vv \rangle$ of translated versus actual neuronal configurations. In both directions and for all pairs ($T \to S$ and $S \to T$), translated configurations preserve mean activity with high fidelity, and captured substantial pairwise covariances, except for a few outliers discussed bellow (see Fig. 4 B-C). These values were comparable to, and in some cases better than, within-animal reconstructions, indicating that translation via the latent space preserved meaningful structure. A detailed examination of these outliers reveals that the model fails to reproduce the mean activity $\langle v_j \rangle$ of 2% to 18% of their neurons (see Fig. S9 A-B). These neurons tend to be spatially structured like functional ensembles (see Fig. S9 C), and we show in Fig. S9 D that this is consistently due to $< 5$ hidden units, which apparently failed to maintain the teacher's spatial distribution. It is possible that by re-training those outlier student RBMs, the failing hidden units would retain teacher characteristics and therefore remove this problem. However, these examples illustrate how a few hidden units can have a significant detrimental impact, confirming the robustness of the method the rest of the time.

**Translated patterns are probable under the receiving model.** To test whether translated configurations were plausible under the receiving RBM, we evaluated their free energy $F(\mathbf{v})$, which measures the log-probability of the configurations up to an additive constant. Translated patterns ($T \to S$ and $S \to T$) had free energies comparable to those of within-fish reconstructions ($T \to T$, $S \to S$), and well within the expected distribution for each fish (Fig. 4 D). In contrast, randomized activity patterns yield, under the same reconstruction procedure, significantly higher free energy and thus

are unlikely under the receiving model. This latter test demonstrates that the translation accuracy is not a simple byproduct of the mapping procedure, but reflects the accuracy of the representation in hidden space. These results held across nearly all teacher–student pairs, with the exception of the outliers mentioned above (Fig. 4 E).

**Spatial structure is retained across fish.** We finally asked whether the transferred activity patterns retained the spatial features of the source configuration. We computed the spatial correlation between the original and translated configurations at all time points. With the exception of the outliers described above (see Fig. S10), both $T \to S$ and $S \to T$ translations achieved significantly higher spatial correlations (median $\approx 0.55$) than time-shuffled controls (median $\approx 0.3$; Fig. 4 F–G), indicating that large-scale spatial motifs are preserved during transfer.

These results demonstrate that spontaneous neuronal activity patterns can be robustly translated across individuals through the shared RBM latent space. These translated configurations preserve key statistical and spatial features, are assigned high probability by the recipient model, and closely resemble the native activity observed in each fish. These findings indicate that the learned latent space encodes a shared and stereotyped repertoire of spontaneous brain-wide activity motifs. This framework offers a powerful new approach for comparing high-dimensional neuronal dynamics across animals, even in the absence of cell-by-cell correspondence.

# Discussion

In this study, we uncovered a shared latent structure underlying spontaneous brain-wide activity in larval zebrafish. By using Latent-aligned Restricted Boltzmann Machines (LaRBMs) across individuals, we identified a common latent space composed of functional assemblies: hidden units that represent recurrent co-activation patterns. The activity of one animal could be projected into this space and transferred into the neuronal space of another, preserving spatial structure and probabilistic consistency. These results reveal that spontaneous activity is structured by a shared repertoire of latent cell assemblies, suggesting conserved circuit-level constraints across brains.

Even in the absence of constraints that enforce a shared latent space, many co-activation motifs were spontaneously conserved across individuals. When RBMs were trained independently on individual fish, approximately 30% of the hidden units could be quantitatively matched across all animals based on spatial similarity. Visual inspection revealed that up to 50% displayed a conserved spatial density core, likely reflecting functional assemblies shaped by developmental or anatomical regularities. Although LaRBM does not uncover additional motifs beyond those of the teacher, it demonstrates that the teacher's latent space acts as a shared non-linear functional basis, capable of representing spontaneous activity across diverse individuals. The success of this transfer supports the presence of stereotyped latent population codes that generalize beyond individual variability.

The RBMs' shared hidden units exhibited spatial and functional consistency across individuals. This is remarkable because although student models were initialized using the spatial weight distribution of the teacher's RBM, they were trained solely on functional activity without any spatial priors. However, hidden units retained consistent spatial patterns and, in particular, pairwise correlations between hidden units were partially preserved across teacher and student. This correlation structure was not explicitly enforced during training, suggesting that the RBM uncovers latent activity motifs that recur across individuals. These motifs are functionally consistent, as

reflected in preserved hidden-unit correlations, but also spatially conserved, despite the absence of explicit spatial constraints.

Moreover, the fact that the transferred activity patterns preserve low free energy in the recipient model indicates that the latent space captures not only structural regularities but also statistical plausibility, an essential property for generative models. Analogously to Helmholtz free energy in statistical physics, the free energy of an activity pattern measures the cost of maintaining a particular neural configuration (macrostate) by marginalizing over all compatible hidden configurations (microstates). Low free energy implies high probability under the model. This framework enables cross-individual comparison of brain-wide activity based on statistically grounded latent structure, allowing for quantitative evaluation of functional similarity grounded in principles from statistical physics.

Compared to other approaches for inter-subject alignment, such as anatomical registration, shared response modeling (SRM), hyperalignment, or contrastive autoencoders, our method offers several advantages. It is generative, allowing brain-wide activity patterns to be sampled, evaluated, and transferred probabilistically. It defines a probabilistic prior over spontaneous neural configurations, and after latent-alignment, this prior is shared across individuals, reflecting common structure in brain-wide activity. It is compositional, combining latent assemblies non-linearly to capture a richer repertoire of motifs than possible in linear subspaces. And it retains single-cell interpretability, unlike deep autoencoder methods. In addition, our approach is fully unsupervised and does not require stimuli, behavioral labels, or neuron correspondence.

Despite these strengths, the current implementation of our approach has limitations. It assumes, on the mesoscal, a spatial organization of brain function. Although this assumption is appropriate for the larval zebrafish brains at 20–30 $\mu$m resolution, it may not generalize to a more variable or less stereotyped system, such as the mammalian cortex. Furthermore, the need to select a reference (teacher) fish introduces a source of bias. Although we observed robust generalization across multiple teacher choices, future extensions could mitigate this issue by implementing joint training or multi-teacher constraints. A symmetric, group-constrained approach, where all models are trained jointly with shared constraints, could enable the discovery of a richer and more balanced repertoire of conserved motifs. This would avoid teacher-specific bias and improve the statistical power to detect rare or subtle shared features. The resulting latent space could serve as a functional atlas onto which new data could be projected and compared across developmental stages, genotypes, or experimental paradigms. This framework could be used for functional phenotyping, where an individual's deviation from the typical use of latent space may indicate neurodevelopmental or pathological alterations.

The emergence of spatially and statistically conserved latent motifs suggests that spontaneous dynamics are not arbitrary, but shaped by shared developmental programs and circuit architecture. The ability to identify and compare these motifs across individuals, without relying on anatomical correspondence or external stimuli, provides a new lens on how vertebrate brains are organized and how they differ. Together, our findings lay the foundations for a comparative neuroscience framework based on shared latent structure in spontaneous activity, enabling quantitative, model-based comparisons of brain-wide dynamics across individuals, developmental stages, or experimental conditions.

# Materials and Methods

## Data and Code Availability

All code, models, and post-processed data used in the present article are available from a centralized repository at `https://github.com/EmeEmu/ZebrafishSharedRBMs`. Datasets and models are available for download at `https://doi.org/10.5281/zenodo.17121390`. Raw data is available upon request (approximately 400GB).

Calcium imaging pre-processing was performed using previously published protocols and software implemented in MATLAB (Mathworks) [31, 43, 26, 38].

Morphological Registrations was performed using the Advanced Normalization Tools (ANTs) [39].

The Restricted Boltzmann Machine modeling is based on the general-purpose Julia package `https://github.com/cossio/RestrictedBoltzmannMachines.jl` [11]. An extension to this package implementing the LaRBM procedure is available at `https://github.com/BiRBMs2024/BiTrainedRBMs.jl/tree/main`.

## Fish Husbandry

All experiments were performed on nacre Tg(elavl3:H2B-GCaMP6f) Danio *rerio* larvae aged 5 to 6 days post-fertilization (dpf). Larvae were reared in groups of 20-30, in Petri dishes in embryo medium (E3) on a 14/10 hour light/dark cycle at $28°C$, and were fed $1mL$ of rotifers at $10^6/\ell$ daily from 5 dpf..

Larvae were screened at 4 and 5dpf for GCaMP6f expression, checking for fluorescence intensity and uniform expression patterns. At 5dpf, larvae were also screened for correct inflation of the swim bladder.

The experimental protocols were approved by Le Comitée d'Ethique pour l'Expérimentation Animale Charles Darwin C2EA-05 (02601.01 and #32423-202107121527185 v3).

## Experimental Protocol

### Fish Preparation

Larvae were paralyzed by immersion in a $1mg/mL$ solution of alpha-bungarotoxin (Invitrogen™ B1601) for one to two minutes in groups of 5 fish. After treatment, the larvae were washed and placed in E3 solution for 20 minutes. Startle response and blood flow were then checked to assess correct paralysis.

A single larvae is then mounted tail first in a capillary (Wiretrol® II $25-50\mu L$) in 2% low melting point agarose (Invitrogen™ 16520-100), and placed in the imaging tank, in E3 solution, under the microscope for an habituation period of 1 hour.

### Imaging

Spontaneous neural activity was recorded using Light-Sheet Microscopy, at 2.5 volumes per second during 25 minutes, with 25 z-planes separated by $10\mu m$ [31].

A total of 17 fish were recorded. Out of those, 11 were discarded : 3 due to failing paralysis, 2 died during recording, 4 presented uncorrectable drift in z, and 2 presented very low fluorescence signal.

This leaves 6 fish which were pre-processed and used in the present study. Fish 1 and 3 were randomly chosen as examples in all main figures, with Fish 3 as teacher and Fish 1 as student in all relevant figures (see Tab. 1).

| Fish # | Clutch | Session | Age (dpf) | Time of day | Neurons |
|---|---|---|---|---|---|
| **1** | A | 1 | 6 | 14:30 | 44378 |
| 2 | A | 1 | 6 | 17:30 | 48179 |
| **3** | A | 1 | 6 | 13:00 | 43379 |
| 4 | B | 2 | 5 | 16:30 | 41349 |
| 5 | B | 2 | 5 | 17:45 | 40242 |
| 6 | B | 3 | 6 | 12:00 | 43354 |

**Table 1.** List of fish

### Data pre-processing

Image pre-processing was performed offline using MATLAB according to previously published protocols [31, 43, 26].

Two dimensional drift correction was applied to each imaging layer to correct for movements of the sample during recording. Watershed was then used for cell segmentation, and a mean fluorescence value was computed from each cell ROI $i$ and frame $t$ : $F_{it}$. This was then normalized to $\Delta F/F_{it} = (F_i - \langle F_i \rangle_t)/(\langle F_i \rangle_t - F_0)$, where $\langle F \rangle_t$ is the baseline fluorescence of neuron $i$ and $F_0$ is the image-wide background fluorescence.

The $\Delta F/F_{it}$ was then deconvolved into binarized spike trains $s_{it} \in [0, 1]$ using Blind Sparse Deconvolution (BSD) [38], using the calcium kernel time constants for GCaMP6f previously inferred using BSD [26, 41] : rise time 0.15s and decay time 1.6s. This method was used over other published methods as it not only deconvolves the fluorescence into spike trains, but also provides a rationale for the binarization of this train.

The first 4 minutes (600 frames) of each recording were excluded. The fish were accustomed to the imaging chamber for 1 hour with the laser turned off to avoid photobleaching. When the laser is turned on at the start of recording, the mean calcium signal rises sharply and then decays to a stable level within about 1 min. By discarding the initial 4 min, we ensure that all analyzed data come from the steady-state phase of neuronal activity under continuous illumination.

Samples of the binarized whole-brain activity for each fish can be seen in Movie S1.

### Morphological Registration

For each recording, a high-resolution stack was acquired with 250 z planes separated by $1\mu m$ and a camera exposure time of $200ms$ per layer. A low-resolution stack was then created by averaging all frames of the functional recording, and then rigidly registered to the high-resolution stack using the Advanced Normalization Tools (ANTs) [39].

The high-resolution stacks of all fish were then registered together using both affine and warp transformations to create a single reference space in which all segmented neurons could be projected. The ANTs command and parameters were adapted from Vladimirov et al. [42], and are given bellow.

```
antsRegistration
    --float 1 --dimensionality 3 --interpolation WelchWindowedSinc --use-histogram-matching 0
    --output [outputdir/LowRes_TO_HighRes_ ,outputdir/LowRes_TO_HighRes.nrrd]
    --initial-moving-transform [HighRes.nrrd,LowRes.nrrd,1]
    --transform Rigid[0.1] --metric MI[HighRes.nrrd,LowRes.nrrd,1,32,Regular,0.25]
    --convergence [200x200x200x0,1e-8,10] --shrink-factors 12x8x4x2 --smoothing-sigmas 4x3x2x1vox
    --transform Affine[0.1] --metric MI[HighRes.nrrd,LowRes.nrrd,1,32,Regular,0.25]
    --convergence [200x200x200x0,1e-8,10] --shrink-factors 12x8x4x2 --smoothing-sigmas 4x3x2x1vox
```

```
antsRegistration
    --float 1 --dimensionality 3 --interpolation WelchWindowedSinc --use-histogram-matching 0
    --output [HighRes_TO_Reference_ ,HighRes_TO_Reference.nrrd]
    --initial-moving-transform [Reference.nrrd,HighRes.nrrd,1]
```

```
    --transform Rigid[0.1] --metric MI[Reference.nrrd,HighRes.nrrd,1,32,Regular,0.25]
    --convergence [200x200x200x0,1e-8,10] --shrink-factors 12x8x4x2 --smoothing-sigmas 4x3x2x1vox
    --transform Affine[0.1] --metric MI[Reference.nrrd,HighRes.nrrd,1,32,Regular,0.25]
    --convergence [200x200x200x0,1e-8,10] --shrink-factors 12x8x4x2 --smoothing-sigmas 4x3x2x1vox
    --transform SyN[0.05,6,0.5] --metric CC[Reference.nrrd,HighRes.nrrd,1,2]
    --convergence [200x200x200x200x10,1e-7,10] --shrink-factors 12x8x4x2x1
    --smoothing-sigmas 4x3x2x1x0vox


antsApplyTransforms
    --dimensionality 3 --float --input LowRes.nrrd --reference-image Reference.nrrd
    --output outputdir/LowRes_TO_Reference.nrrd --interpolation WelchWindowedSinc --default-value 0
```

## Restricted Boltzmann Machines

### Definition

Restricted Boltzmann machines (RBM) [18] are two-layer energy-based models, over $N$ visible units $\mathbf{v} = (v_1, \ldots, v_N)$ representing the neural activity with binary values $v_i \in \{0, 1\}$, and $M$ real-valued hidden units $\mathbf{h} = (h_1, \ldots, h_M)$.

The RBM defines a probability distribution over configurations of all the units given by:

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v}, \mathbf{h})} \tag{3}$$

where $Z$ is a normalization constant known as the partition function, and $E(\mathbf{v}, \mathbf{h})$, the energy, is given by:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^{N} \mathcal{V}_i(v_i) + \sum_{\mu=1}^{M} \mathcal{U}_\mu(h_\mu) - \sum_{i=1}^{N} \sum_{\mu=1}^{M} w_{i\mu} v_i h_\mu \tag{4}$$

The functions $\mathcal{V}_i(v_i)$ and $\mathcal{U}_\mu(h_\mu)$ are potentials biasing the activities of single units, while the weights $w_{i\mu}$ account for interactions between visible and hidden units. In the case of neuronal data, we take $\mathcal{V}_i(v_i) = g_i v_i$ with $g_i$ called the field. Following van der Plas et al. [41], we use dReLU potentials for the hidden units,

$$\mathcal{U}_\mu(h_\mu) = \frac{1}{2} \frac{\gamma_\mu}{1 + \eta_\mu} h_{\mu,+}^2 + \frac{1}{2} \frac{\gamma_\mu}{1 - \eta_\mu} h_{\mu,-}^2 + \theta_\mu h_\mu + \frac{\Delta_\mu}{1 + \eta_\mu} h_{\mu,+} - \frac{\Delta_\mu}{1 - \eta_\mu} h_{\mu,-} \tag{5}$$

with real parameters $\gamma_\mu, \eta_\mu, \Delta_\mu, \theta_\mu$, satisfying $\gamma_\mu > 0$. Here $h_{\mu,+} = \max(h_\mu, 0)$ and $h_{\mu,-} = \min(h_\mu, 0)$.

To obtain the probability of neural activity, the hidden units are marginalized. In this way, one defines a marginal likelihood over the visible units:

$$P(\mathbf{v}) = \int P(\mathbf{v}, \mathbf{h}) \mathrm{d}\mathbf{h} = \frac{1}{Z} e^{-F(\mathbf{v})} \tag{6}$$

where $F(\mathbf{v})$ is a free energy function that incorporates the effective interactions induced by the marginalized hidden variables,

$$F(\mathbf{v}) = \sum_{i=1}^{N} \mathcal{V}_i(v_i) - \sum_{\mu=1}^{M} \ln \int e^{\sum_i w_{i\mu} v_i h_\mu - \mathcal{U}_\mu(h_\mu)} dh_\mu \tag{7}$$

### Standardized RBMs

Following [12], we employ a generalized *centering trick* [28], where the activity of the hidden units is not only centered, but also standardized.

The Standardized Restricted Boltzmann machine (StdRBM) is a re-parameterization of the RBM, where the energy function is defined as follows:

$$E(\mathbf{v}, \mathbf{h}) = \cdots - \sum_i g_i v_i - \sum_\mu \theta_\mu h_\mu - \sum_{i\mu} w_{i\mu} \frac{v_i - \lambda_i}{\sigma_i} \frac{h_\mu - \lambda_\mu}{\sigma_\mu} + \sum_{i\mu} \frac{w_{i\mu}}{\sigma_i \sigma_\mu} \lambda_i \lambda_\mu \tag{8}$$

where we only show the fields for the unit potentials for simplicity. During training, the parameters $\lambda_i, \lambda_\mu$ and $\sigma_i, \sigma_\mu$ track a moving average of the mean and standard deviations of the corresponding unit activities. After training, the StdRBM can be converted into an equivalent RBM by the transformations:

$$g_i \to g_i - \sum_\mu \frac{w_{i\mu}}{\sigma_i \sigma_\mu} \lambda_\mu, \quad \theta_\mu \to \theta_\mu - \sum_i \frac{w_{i\mu}}{\sigma_i \sigma_\mu} \lambda_i, \quad w_{i\mu} \to \frac{w_{i\mu}}{\sigma_i \sigma_\mu} \tag{9}$$

As has been observed empirically in previous works [28, 12, 25], such centering and standardization of unit activities leads to improved training convergence and stability.

### Training

All RBM parameters $(g_i, \gamma, \eta, \Delta, \theta, w_{i\mu})$ are trained to maximize the likelihood of a neural activity dataset. If a data set of neural activity recordings is given as $\mathcal{D} = \{\mathbf{v}^t\}$, where $t$ is the index of the sample (or time), then the RBM is trained by maximizing:

$$\frac{1}{|\mathcal{D}|} \sum_{t \in \mathcal{D}} \ln P(\mathbf{v}_t) - \mathcal{R}(W) \tag{10}$$

Here, $\mathcal{R}(W)$ is a regularization term applied over the RBM weights. Following van der Plas et al. [41], we employ an $L_1$ regularization for single-cell RBMs,

$$\mathcal{R}(W) = \lambda_1 \sum_{i\mu} |w_{i\mu}| \tag{11}$$

where $\lambda_1$ controls the regularization strength. As in van der Plas et al. [41], we set $\lambda_1 = 0.02$, and $M = 100$ hidden units. The optimization is carried out by gradient ascent of Eq. (10). We use the ADAM algorithm [22].

### Sampling and Reconstruction

Once trained, the RBM can generate synthetic neuronal configurations by alternating between sampling hidden states from $P(h \mid v)$ and visible states from $P(v \mid h)$, forming a Gibbs sampling chain. This process allows the model to reconstruct expected activity patterns from observed configurations and explore the distribution it has learned. Movie S2 illustrates a single Gibbs sampling step and the resulting reconstruction $\langle v \rangle$ from an observed configuration in Fish 1.

### Evaluation

Trained RBMs are evaluated on their ability to reproduce data statistics. As they are trained to maximize the log-likelihood of the data, a well trained RBM is expected to generate data with statistics $\langle v_i \rangle$, $\langle h_\mu \rangle$, and $\langle v_i h_\mu \rangle$ matching the corresponding empirical data statistics. We also evaluate the RBMs' ability to reproduce the second-order statistics $\langle v_i v_j \rangle - \langle v_i \rangle \langle v_j \rangle$ and $\langle h_\mu h_\nu \rangle - \langle h_\mu \rangle \langle h_\nu \rangle$ to ensure that first order interactions between neurons are captured by the model [41]. An RBM has to reproduce well all five statistics to be considered usable in our use case.

Measurement of statistics matching between model-generated and empirical data is done using the normalized Root Mean Squared Error (nRMSE) as introduced by van der Plas et al. [41] (see Measuring identity). This measurement is standardized such that a value of 1 corresponds to shuffled data statistics, and a value of 0 corresponds to the optimal matching expected from the difference between train and test sets. The evaluation of an RBM is thus represented by a vector of length 5 containing the nRMSE for all five statistics, and multiple RBMs can be compared using the $L_\infty$ norm.

### Inferred neuronal couplings

van der Plas et al. [41] showed that an effective coupling matrix $J_{ij}$ could be inferred from a trained RBM model. This is done by manually perturbing the activity of each neuron and quantifying the impact on other neurons by evaluating the marginal distribution $P(\mathbf{v})$. Given a neuronal configuration $\mathbf{v}$, the coupling is then defined as

$$\begin{aligned} J_{ij}(\mathbf{v}) = \log &\left( \frac{P\left(v_i = 1 \mid v_1, \cdots, v_j = 1, \cdots v_N\right)}{P\left(v_i = 1 \mid v_1, \cdots, v_j = 0, \cdots v_N\right)} \right) \\ - \log &\left( \frac{P\left(v_i = 0 \mid v_1, \cdots, v_j = 1, \cdots v_N\right)}{P\left(v_i = 0 \mid v_1, \cdots, v_j = 0, \cdots v_N\right)} \right) \end{aligned} \tag{12}$$

The overall coupling matrix is then obtained from all empirical neuronal configurations as :

$$J_{ij} = \langle J_{ij}(\mathbf{v}) \rangle_{\text{data}} \tag{13}$$

However, calculating $J_{ij}(\mathbf{v})$ is computationally demanding, and we use the following approximation introduced by van der Plas et al. [41]:

$$J_{ij} \approx \sum_{\mu=1}^{M} w_{i\mu} w_{j\mu} \langle \text{Var}(h_\mu \mid \mathbf{v}) \rangle_{\text{data}} \tag{14}$$

## Voxelised RBMs

### Voxelisation

To obtain a coarse-grained description of the datasets, we voxelize the brain into cubes of side $v$ (typically $v = 20\mu m$). This grid is defined in the common registered brain space. Neurons are assigned to the voxels after segmentation, according to their center of mass. The activity of each voxel for each fish is then defined as :

$$v_m(t) = \frac{1}{|\mathcal{V}_m|} \sum_{i=1}^{|\mathcal{V}_m|} \frac{\Delta F}{F}_i(t) \tag{15}$$

where $v_m$ is the activity of voxel $m$ containing the set of neurons $\mathcal{V}_m$ and $\frac{\Delta F}{F}_i(t)$ is the normalized fluorescence signal of neuron $i$ at time $t$.

This activity is then normalized for each fish $f$ and each voxel using z-score :

$$\mathrm{v}_m^f(t) = \frac{v_m^f(t) - \langle v_m^f \rangle_t}{\sigma_t(v_m^f)} \tag{16}$$

This process yields a series of voxelized activities $\mathrm{v}_m^f(t)$ where each voxel $m$ corresponds to the same anatomical position across all fish.

To ensure that the z-scored activity of each voxel is well defined, we discard voxels for which $|\mathcal{V}_m^f| < 2$ for at least one fish.

### Training

Contrary to the neuronal RBMs described above (Restricted Boltzmann Machines), in the case of voxelized data, we used Standardized RBMs with gaussian visible units, as the voxelized activity approximately follows a normal distribution (see Fig. S3 C).

$$\mathcal{U}_i(v) = \frac{1}{2}\gamma_i v^2 + \theta_i v \tag{17}$$

We found empirically that for voxelized RBMs, $L_1$ regularization produced weight matrices where the majority of hidden units were decoupled from the visible layer. Hence we use the $L_1^2$ regularization introduced by Tubiana, Cocco, and Monasson [37] :

$$\mathcal{R}(W) = \frac{\lambda_{21}}{N} \sum_{\mu} \left( \sum_{i} |w_{i\mu}| \right)^2 \tag{18}$$

where $N$ is the number of voxels and $\lambda_{21}$ is the regularization factor.

We performed a cross-validation over the number of hidden units $M$ and regularization factor $\lambda_{21}$ on Fish 1 (Fig. S3 A). We identified 3 sets of hyperparameters which we then tested on the concatenated voxelized data of all 6 fish in the dataset (Fig. S3 B), and chose $M = 40$ and $\lambda_{21} = 0.1$.

The rest of the training parameters are listed here :

- 20000 gradient updates.
- 50 Markov Chain steps between gradient updates.
- 100 time points per batch.
- $5 \cdot 10^{-4}$ learning rate for the first 1/4 of the training. The learning rate is then annealed with a geometrical decay to reach $10^{-5}$ at the end of training.

## Latent-aligned Restricted Boltzmann machines (LaRBMs)

### Teacher RBMs

Teacher standardized RBMs were constructed and trained from whole-brain binarized neuronal activity with the following parameters :

- $M = 100$ hidden units (instead of either 200 or 100 in van der Plas et al. [41] as most of our fish have weak activity in the optic tectum).
- $\lambda_1 = 0.02$ for weight L1 regularization (as obtained by van der Plas et al. [41] after cross-validation).
- 200000 gradient updates.
- 15 Markov Chain steps between gradient updates.
- 256 time points per batch.
- $5 \cdot 10^{-4}$ learning rate for the first 1/4 of the training. The learning rate is then annealed with a geometrical decay to reach $10^{-5}$ at the end of training.

Before training, individual fish datasets were split into a 70% training and 30% validation sets using the method described by van der Plas et al. [41]. The validation set was used to evaluate RBM-generated vs empirical data statistics as described in Evaluation.

For every fish, we trained 10 RBMs and kept the one which performed best as described in Evaluation (see Fig. S5 for all training statistics). Indeed, we found that convergence can be very variable, with some models converging satisfactorily, and other either converging poorly or not converging at all.

### Student LaRBM Initialization

Student LaRBMs are initialized from the teacher with the following steps.

First, the fields $g_j$ of the visible layer are estimated from logit of the average neuron activity $\langle v_j \rangle$ : $g_j = -\log\left(\frac{1}{\langle v_j \rangle} - 1\right)$. The visible scaling ($\sigma_j$) and bias ($\lambda_j$) parameters used to standardize the RBM are also estimated from neuronal activity as : $\sigma_j = \frac{1}{2}\,\mathrm{var}(v_j) + 1$ and $\lambda_j = \langle v_j \rangle$.

Second, hidden potentials $\mathcal{U}_\mu(h_\mu)$ (i.e. all parameters of the dReLU $\gamma_\mu, \eta_\mu, \Delta_\mu, \theta_\mu$) are copied directly from the teacher RBM.

Third, the teacher weight matrix $w_{i\mu}$ is spatially interpolated at the locations of student neurons. To do so, we first build a 3D map of weights for each hidden unit $\mu$ :

$$W_\mu(\vec{x}) = \frac{\sum_i w_{i\mu}\delta(\vec{x}, \vec{x}_i)}{\sum_i \delta(\vec{x}, \vec{x}_i)} \tag{19}$$

$$\delta(\vec{x}, \vec{y}) = \begin{cases} 1, & \text{if } ||\vec{x} - \vec{y}|| < r \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

where $\vec{x}$ is any 3D location in shared brain space, $\vec{x}_i$ is the 3D location of teacher neuron $i$, and $\delta(\vec{x}, \vec{x}_i)$ is a boolean ball of radius $r = 4\mu m$, the average size of neurons. These maps are then convolved with a gaussian kernel $G(\vec{x}, \sigma)$ of width $\sigma = 4\mu m$ to smooth out local discontinuities : $\mathcal{W}_\mu(\vec{x}) = W_\mu(\vec{x}) * G(\vec{x}, \sigma)$. Because weight matrices are very sparse, this smoothing step has a tendency to attenuate the weight map. We therefore add a bias :

$$\beta_\mu = \min_\beta \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(w_{i\mu} - \beta\mathcal{W}_\mu(\vec{x_i})\right)^2} \tag{21}$$

Finally, a weight matrix for the student can be created by evaluating $\mathcal{W}_\mu(\vec{x})$ at the locations $\vec{x}_j$ of neurons from the student fish :

$$w_{j,\mu} = \beta_\mu \mathcal{W}_\mu(\vec{x}_j) \tag{22}$$

### Student training

Once initialized, student RBMs are trained with the same set of parameters as the teacher, except for the number of gradient updates for which we found 1/10th (20000) to be sufficient for convergence.

In absence of any constraints, the space of representations learned by the student can drift arbitrarily far from the one of the teacher. This is undesirable because we cannot compare the hidden unit activities from the teacher and the student. In order to avoid this, we introduce a new training paradigm to maintain an alignment between the hidden units of the student and the teacher. Mathematically, we would like the marginal distribution of hidden unit activities, $P(\mathbf{h}) = \sum_{\mathbf{v}} P(\mathbf{v}, \mathbf{h})$, to be similar in the teacher and the student. To achieve this, we first sample hidden unit activity from the teacher RBM to create a dataset of hidden unit activities, $\mathcal{T} = \{\mathbf{h}_t\}$ from the teacher. The student is trained following the modified objective:

$$\frac{1}{|\mathcal{D}|}\sum_s \ln P(\mathbf{v}_s) + \frac{\kappa}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \ln P(\mathbf{h}_t) - \mathcal{R}(W) \tag{23}$$

The first and last terms are the same appearing in the original training objective, Eq. (10). The second term is the average log-likelihood of the teacher hidden unit activity evaluated under the student. By maximizing this term, we force the student to align its

marginal hidden unit distribution to match that of the teacher. The coefficient $\kappa$ controls the relative importance of this term: for $\kappa = 0$, we recover the original training objective of Eq. (10), while for $\kappa$ large, the student RBM is forced to match the teacher hidden unit statistics more closely.

### Mapping from one fish to another

In principle, translating activity from the visible layer of one RBM, through the hidden layer, to the visible layer of another RBM should be done via sampling the hidden layer given the first fish visible layer, and for each sample, sampling the second fish visible layer given the hidden layer, and then averaging over all samples, which yields $\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\mathbb{E}_{\mathbf{v}^S|\mathbf{h}^T}\left[\mathbf{v}^S\right]$. However, in practice, this is computationally intensive and thus impractical in our use case. Instead, we use $\mathbb{E}_{\mathbf{v}^S|\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}}\left[\mathbf{v}^S\right]$ as we have found empirically that for a high number ($\sim 1000$) of samples the two method give similar results (see Fig. S8 A-D).

This similarity can be explained with a Taylor expansion on $f_i : \mathbf{h}^T \mapsto \mathbb{E}_{\mathbf{v}^S|\mathbf{h}^T}\left[\mathbf{v}_i^S\right]$ at $\mathbf{h}^{T*} = \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]$:

$$\begin{aligned} f_i\left(\mathbf{h}^T\right) &\underset{\mathbf{h}^T \to \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}[\mathbf{h}^T]}{=} f_i\left(\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]\right) \\ &+ \sum_\mu \left(\mathbf{h}_\mu^T - \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}_\mu^T\right]\right) \frac{\partial f_i}{\partial \mathbf{h}_\mu^T}\left(\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]\right) \\ &+ \frac{1}{2}\sum_{\mu,\nu} \left(\mathbf{h}_\mu^T - \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}_\mu^T\right]\right)\left(\mathbf{h}_\nu^T - \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}_\nu^T\right]\right) \frac{\partial^2 f_i}{\partial \mathbf{h}_\mu^T \partial \mathbf{h}_\nu^T}\left(\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]\right) \\ &+ \dots \end{aligned} \tag{24}$$

Because of the conditional independence of $\mathbf{h}^T$ given $\mathbf{v}^T$,

$$\begin{aligned} \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[f_i\left(\mathbf{h}^T\right)\right] &\underset{\mathbf{h}^T \to \mathbf{h}^{T*}}{=} f_i\left(\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]\right) \\ &+ \frac{1}{2}\sum_\mu \underset{\mathbf{h}^T|\mathbf{v}^T}{\mathrm{var}}\left[\mathbf{h}_\mu^T\right] \frac{\partial^2 f_i}{\partial \mathbf{h}_\mu^{T2}}\left(\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]\right) \\ &+ \dots \end{aligned} \tag{25}$$

Since $f_i\left(\mathbf{h}^T\right) = \mathbb{E}_{\mathbf{v}^S|\mathbf{h}^T}\left[\mathbf{v}_i^S\right] = \sigma\left(g_i + \sum_\mu w_{i\mu}^S \mathbf{h}_\mu^T\right)$, where $\sigma(z) = \frac{1}{1+\exp(-z)}$, we find

$$\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\mathbb{E}_{\mathbf{v}^S|\mathbf{h}^T}\left[\mathbf{v}^S\right] = \mathbb{E}_{\mathbf{v}^S|\mathbf{h}^{T*}}\left[\mathbf{v}^S\right] + \Delta \tag{26}$$

where

$$\Delta_i \approx \frac{1}{2}\sum_\mu \underset{\mathbf{h}^T|\mathbf{v}^T}{\mathrm{var}}\left[\mathbf{h}_\mu^T\right]\left(w_{i\mu}^S\right)^2 \sigma''\left(g_i + \sum_\mu w_{i\mu}^S \mathbf{h}_\mu^T\right) \tag{27}$$

We show on Fig. S8 that $\Delta$ is indeed negligible compared to $\mathbb{E}_{\mathbf{v}^S|\mathbf{h}^{T*}}\left[\mathbf{v}^S\right]$.

## Measuring identity

As introduced by van der Plas et al. [41], we quantify the goodness of fit to the identity using the normalized Root Mean Squared Error (nRMSE). It is defined as :

$$\mathrm{nRMSE}(X, Y) = 1 - \frac{\mathrm{RMSE}(X, Y) - \mathrm{RMSE}(\Pi(X), \Pi(Y))}{\mathrm{RMSE}(X^O, Y^O) - \mathrm{RMSE}(\Pi(X), \Pi(Y))} \tag{28}$$

where $X$ and $Y$ are two vectors of same length, $\mathrm{RMSE}(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(A_i - B_i\right)^2}$ the standard Root Mean Squared Error, $(X^O, Y^O)$ is the pair of optimal vectors corresponding to the best expected fit of $X$ and $Y$, and $\Pi$ is a shuffling operator such that $\Pi(A)_i = A_{\pi(i)}$.

Unlike the standard RMSE, this measure is independent of the length of the vectors $X$ and $Y$, and is easily interpretable. Indeed, a value of $\mathrm{nRMSE}(X,Y) = 1$ means that $X$ and $Y$ are as uncorrelated as shuffled vectors, while $\mathrm{nRMSE}(X,Y) = 0$ means that $X$ and $Y$ are as identical as can be expected from the optimal $(X^O, Y^O)$.

In the case where no optimal $(X^O, Y^O)$ can be defined other than $X^O = Y^O$, $\mathrm{RMSE}(X^O, Y^O) = 0$ and therefore :

$$\mathrm{nRMSE}(X,Y) = \frac{\mathrm{RMSE}(X,Y)}{\mathrm{RMSE}(\Pi(X), \Pi(Y))} \tag{29}$$

$\mathrm{nRMSE}(X,Y) = 0$ is then interpreted as $X_i = Y_i \ \forall i$.

## Measuring Spatial Similarity

To compare how a spatially defined variable (such as neuron weight or firing rate) is distributed across two different fish, we define a spatial similarity measure between animals A and B.

Let $\phi^A$ and $\phi^B$ be two spatial fields that assign scalar values $y_i^A \in \mathbb{R}$ and $y_j^B \in \mathbb{R}$ to neurons at 3D positions $\mathbf{x}_i^A$ and $\mathbf{x}_j^B$ in fish A and B, respectively: $\phi^A(\mathbf{x}_i^A) = y_i^A$ and $\phi^B(\mathbf{x}_i^B) = y_i^B$. We wish to estimate how well $\phi^A$ aligns with $\phi^B$.

Since neurons are not spatially matched between individuals, we interpolate the values $y_i^B$ from fish B to the locations of neurons in fish A using Gaussian kernel smoothing :

$$\hat{y}_j^{A\to B} = \frac{\sum_i y_i^A e^{-\frac{\|\mathbf{x}_j^B - \mathbf{x}_i^A\|^2}{2\sigma^2}}}{\sum_i e^{-\frac{\|\mathbf{x}_j^B - \mathbf{x}_i^A\|^2}{2\sigma^2}}} \tag{30}$$

where $\|\mathbf{x}_j^B - \mathbf{x}_i^A\|$ is the euclidean distance between neuron $i$ from fish A and neuron $j$ from fish B, and $\sigma$ is the standard width of a gaussian kernel, which we take to be a neuron size ($\sigma = \frac{8\mu m}{2} = 4\mu m$). This provides an estimate of what the spatial distribution of $y$ from fish A would look like in fish B.

We can then evaluate the similarity between the two spatial fields $\phi^A$ and $\phi^B$ by computing the Pearson correlation coefficient between $y_j^B$ and $\hat{y}_j^{A\to B}$ :

$$\rho^{A\to B} = \frac{\mathrm{cov}(y_j^B, \hat{y}_j^{A\to B})}{\sigma_{y_B^2}\sigma_{\hat{y}_j^{A\to B}}} \tag{31}$$

In practice, the Person correlation is used over other identity tests such as the nRMSE (see Methods Measuring identity) because our observables tend to be sparse and the gaussian interpolation will thus naturally lower $\hat{y}$ ($\hat{y}_i^{A\to A} \approx \alpha y_i^A$ with $\alpha < 1$).

In the case of the weight maps presented in Fig. 3 B-F, $\rho^{A\to B}$ is only computed for neurons whose weights are significant ($> 10^{-5}$), as we are only interested in whether student maps are compatible with teachers, and the regularization imposed on weights during RBM training will force a non-spatial linearity in weight distribution.

## Measuring Diagonal Dominance

In Fig. 3 C-D, we ask whether the spatial distribution of hidden units' weights is maintained between teacher and student RBMs, in a way that allows for their

unambiguous pairwise identification. In other words, we ask whether the spatial distribution of $w^S_{i\mu}$ is closer to $w^T_{i\mu}$ than to any other hidden unit $w^T_{i\nu}$ with $\nu \neq \mu$.

To do so we first compute the matrix $\rho^{TS}_{\mu\nu}$ $(\mu, \nu \in (1, ..., M))$ of spatial correlation (see Methods Measuring Spatial Similarity) between the weight map of hidden unit $\mu$ in the teacher RBM, and hidden unit $\nu$ in the student RBM.

Our hypothesis is confirmed not only if each teacher hidden unit $\mu$ is most similar to its corresponding student unit (i.e., if $\arg\max_\nu \rho^{TS}_{\mu\nu} = \mu$), but also if this match is significantly better than any other (i.e., $\rho^{TS}_{\mu\mu} \gg \rho^{TS}_{\mu\nu}$ for all $\nu \neq \mu$).

To test for this systematically, we perform a bootstrap analysis. We start by computing a *distance* between the diagonal and the next largest element of the matrix :

$$d_\mu = \rho^{TS}_{\mu\mu} - \max_{\nu \neq \mu} \rho^{TS}_{\mu\nu} \tag{32}$$

We next compute the same distance for $n = 5000$ random row permutations of the matrix $\rho^{TS}_{\mu\mu}$ :

$$\tilde{d}_\mu = \rho^{TS}_{\pi(\mu)\mu} - \max_{\nu \neq \pi(\mu)} \rho^{TS}_{\pi(\mu)\nu} \tag{33}$$

where $\pi(\mu)$ is a random permutation of $1, ..., M$ (see Fig. S7 C for an example). We can then compute a p-value measuring the probability that $\rho^{TS}_{\mu\mu} \gg \rho^{TS}_{\mu\nu}$ could be explained by random permutations of the matrix :

$$p = \frac{1}{Mn} \sum_{j=1}^{n} \sum_{\mu=1}^{M} \mathbb{1}\{\tilde{d}_\mu > d_\mu\} \tag{34}$$

It is to be noted that, as some hidden units are completely disconnected from the visible layer ($w_{i\mu} < 10^{-5}, \forall i$), some weight maps are undefined. These are therefore omitted from this measurement.

# Acknowledgments

We would like to thank Simona Cocco, Rémi Monasson, and Jérôme Tubiana for helpful discussions and insights. As well as Ans Imran for his help during his internship.

We would also like to thank Abdelkrim Mannioui, Marco Amaral, Edouard Manzoni and the rest of the team at the fish facility for taking care of the animals, as well as Malika Pierrat and Laura Bacerra-Zapata for taking care of the administration and contracts.

This project has received funding from CNRS, Sorbonne University, and the European Research Council (ERC) under the European Union's Horizon 2020 research innovation program grant agreement number 715980. M.DK had a PhD fellowship from the École Doctorale Frontière de l'Innovation en Recherche et Education - Programme Bettencourt, J. FdCD. was supported by the Université PSL, AI Junior Fellow program, G.FB had a PhD fellowship from the iBio program of Sorbonne University.

# References

[1] Misha B. Ahrens et al. "Whole-Brain Functional Imaging at Cellular Resolution Using Light-Sheet Microscopy". In: *Nature Methods* 10.5 (May 2013), pp. 413–420. ISSN: 1548-7105. DOI: 10.1038/nmeth.2434. URL: https://www.nature.com/articles/nmeth.2434 (visited on 07/27/2025).

[2] Adam A. Atanas et al. "Brain-Wide Representations of Behavior Spanning Multiple Timescales and States in C. Elegans". In: *Cell* 186.19 (Sept. 14, 2023), 4134–4151.e31. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2023.07.035. pmid: 37607537. URL: https://www.cell.com/cell/abstract/S0092-8674(23)00850-4 (visited on 07/27/2025).

[3] Connor Brennan and Alexander Proekt. "A Quantitative Model of Conserved Macroscopic Dynamics Predicts Future Motor Commands". In: *eLife* 8 (July 11, 2019). Ed. by Ronald L Calabrese, William S Ryu, and Elizabeth Cropper, e46814. ISSN: 2050-084X. DOI: 10.7554/eLife.46814. URL: https://doi.org/10.7554/eLife.46814 (visited on 04/07/2023).

[4] György Buzsáki and Andreas Draguhn. "Neuronal Oscillations in Cortical Networks". In: *Science* 304.5679 (June 25, 2004), pp. 1926–1929. DOI: 10.1126/science.1099745. URL: https://www.science.org/doi/10.1126/science.1099745 (visited on 07/27/2025).

[5] Edgar Canario, Donna Chen, and Bharat Biswal. "A Review of Resting-State fMRI and Its Use to Examine Psychiatric Disorders". In: *Psychoradiology* 1.1 (May 11, 2021), p. 42. DOI: 10.1093/psyrad/kkab003. pmid: 38665309. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10917160/ (visited on 07/27/2025).

[6] Po-Hsuan Chen et al. *A Convolutional Autoencoder for Multi-Subject fMRI Data Aggregation*. Aug. 17, 2016. DOI: 10.48550/arXiv.1608.04846. arXiv: 1608.04846 [stat]. URL: http://arxiv.org/abs/1608.04846 (visited on 07/27/2025). Pre-published.

[7] Po-Hsuan Chen et al. "A Reduced-Dimension fMRI Shared Response Model". In: *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*. Vol. 1. NIPS'15. Cambridge, MA, USA: MIT Press, Dec. 7, 2015, pp. 460–468.

[8] Simona Cocco et al. "Functional Networks from Inverse Modeling of Neural Population Activity". In: *Current Opinion in Systems Biology* 3 (June 2017), pp. 103–110. ISSN: 24523100. DOI: 10.1016/j.coisb.2017.04.017. URL: https://linkinghub.elsevier.com/retrieve/pii/S2452310017300215 (visited on 08/22/2025).

[9] J. S. Damoiseaux et al. "Consistent Resting-State Networks across Healthy Subjects". In: *Proceedings of the National Academy of Sciences* 103.37 (Sept. 12, 2006), pp. 13848–13853. DOI: 10.1073/pnas.0601417103. URL: https://www.pnas.org/doi/10.1073/pnas.0601417103 (visited on 07/27/2025).

[10] Anastasia Dimakou et al. "The Predictive Nature of Spontaneous Brain Activity across Scales and Species". In: *Neuron* (Mar. 17, 2025). ISSN: 0896-6273. DOI: 10.1016/j.neuron.2025.02.009. URL: https://www.sciencedirect.com/science/article/pii/S0896627325001278 (visited on 04/07/2025).

[11] Jorge Fernandez-de-Cossio-Diaz, Simona Cocco, and Rémi Monasson. "Disentangling Representations in Restricted Boltzmann Machines without Adversaries". In: *Physical Review X* 13.2 (Apr. 5, 2023), p. 021003. DOI: 10.1103/PhysRevX.13.021003. URL: https://link.aps.org/doi/10.1103/PhysRevX.13.021003 (visited on 07/29/2025).

[12] Jorge Fernandez-de-Cossio-Diaz et al. *Designing Molecular RNA Switches with Restricted Boltzmann Machines*. May 12, 2023. DOI: 10.1101/2023.05.10.540155. URL: https://www.biorxiv.org/content/10.1101/2023.05.10.540155v1 (visited on 07/29/2025). Pre-published.

[13] Charles Fieseler et al. *An Intrinsic Neuronal Manifold Underlies Brain-Wide Hierarchical Organization of Behavior in C. Elegans*. Mar. 11, 2025. DOI: 10.1101/2025.03.09.642241. URL: https://www.biorxiv.org/content/10.1101/2025.03.09.642241v1 (visited on 07/27/2025). Pre-published.

[14] Hanna K. Hausman et al. "The Role of Resting-State Network Functional Connectivity in Cognitive Aging". In: *Frontiers in Aging Neuroscience* 12 (June 12, 2020). ISSN: 1663-4365. DOI: 10.3389/fnagi.2020.00177. URL: https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2020.00177/full (visited on 07/27/2025).

[15] James V Haxby et al. "Hyperalignment: Modeling Shared Information Encoded in Idiosyncratic Cortical Topographies". In: *eLife* 9 (June 2, 2020). Ed. by Chris I Baker and Floris P de Lange, e56601. ISSN: 2050-084X. DOI: 10.7554/eLife.56601. URL: https://doi.org/10.7554/eLife.56601 (visited on 07/27/2025).

[16] James V. Haxby et al. "A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex". In: *Neuron* 72.2 (Oct. 20, 2011), pp. 404–416. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2011.08.026. URL: https://www.sciencedirect.com/science/article/pii/S0896627311007811 (visited on 07/27/2025).

[17] G. E. Hinton and R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786 (July 28, 2006), pp. 504–507. DOI: 10.1126/science.1127647. URL: https://www.science.org/doi/10.1126/science.1127647 (visited on 07/27/2025).

[18] Geoffrey E. Hinton. "A Practical Guide to Training Restricted Boltzmann Machines". In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer, 2012, pp. 599–619. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_32. URL: https://doi.org/10.1007/978-3-642-35289-8_32 (visited on 07/29/2025).

[19] R. Devon Hjelm et al. "Restricted Boltzmann Machines for Neuroimaging: An Application in Identifying Intrinsic Networks". In: *NeuroImage* 96 (Aug. 1, 2014), pp. 245–260. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2014.03.048. URL: https://www.sciencedirect.com/science/article/pii/S1053811914002080 (visited on 04/07/2023).

[20] Alexander G. Huth et al. "Natural Speech Reveals the Semantic Maps That Tile Human Cerebral Cortex". In: *Nature* 532.7600 (Apr. 2016), pp. 453–458. ISSN: 1476-4687. DOI: 10.1038/nature17637. URL: https://www.nature.com/articles/nature17637 (visited on 07/27/2025).

[21] Alexander G. Huth et al. *PrAGMATiC: A Probabilistic and Generative Model of Areas Tiling the Cortex*. Apr. 14, 2015. DOI: 10.48550/arXiv.1504.03622. arXiv: 1504.03622 [q-bio]. URL: http://arxiv.org/abs/1504.03622 (visited on 07/27/2025). Pre-published.

[22] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* (Dec. 22, 2014). URL: https://www.semanticscholar.org/paper/Adam%3A-A-Method-for-Stochastic-Optimization-Kingma-Ba/a6cb366736791bcccc5c8639de5a8f9636bf87e8 (visited on 07/29/2025).

[23] Emiliano Marachlian et al. "Principles of Functional Circuit Connectivity: Insights From Spontaneous Activity in the Zebrafish Optic Tectum". In: *Frontiers in Neural Circuits* 12 (June 21, 2018). ISSN: 1662-5110. DOI: 10.3389/fncir.2018.00046. URL: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fncir.2018.00046/full (visited on 07/27/2025).

[24] João C. Marques et al. "Internal State Dynamics Shape Brainwide Activity and Foraging Behaviour". In: *Nature* 577.7789 (7789 Jan. 2020), pp. 239–243. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1858-z. URL: https://www.nature.com/articles/s41586-019-1858-z (visited on 04/07/2023).

[25] Jan Melchior, Asja Fischer, and Laurenz Wiskott. "How to Center Deep Boltzmann Machines". In: *Journal of Machine Learning Research* 17.99 (2016), pp. 1–61. URL: http://jmlr.org/papers/v17/14-237.html.

[26] Geoffrey Migault et al. "Whole-Brain Calcium Imaging during Physiological Vestibular Stimulation in Larval Zebrafish". In: *Current Biology* 28.23 (Dec. 3, 2018), 3723–3735.e6. ISSN: 0960-9822. DOI: 10.1016/j.cub.2018.10.017. URL: https://www.sciencedirect.com/science/article/pii/S0960982218313460 (visited on 04/07/2023).

[27] Sebastian Quiroz Monnens et al. "The Recurrent Temporal Restricted Boltzmann Machine Captures Neural Assembly Dynamics in Whole-Brain Activity". In: *eLife* 13 (Sept. 30, 2024). DOI: 10.7554/eLife.98489.2. URL: https://elifesciences.org/reviewed-preprints/98489 (visited on 07/27/2025).

[28] Gregoire Montavon, Mikio Braun, and Klaus-Robert Muller. "Deep Boltzmann Machines as Feed-Forward Hierarchies". In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, Apr. 21–23, 2012, pp. 798–804. URL: https://proceedings.mlr.press/v22/montavon12.html.

[29] Yu Mu et al. "Glia Accumulate Evidence That Actions Are Futile and Suppress Unsuccessful Behavior". In: *Cell* 178.1 (June 27, 2019), 27–43.e19. ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.05.050. URL: https://www.sciencedirect.com/science/article/pii/S009286741930621X (visited on 04/07/2023).

[30] Eva A. Naumann et al. "From Whole-Brain Data to Functional Circuit Models: The Zebrafish Optomotor Response". In: *Cell* 167.4 (Nov. 3, 2016), 947–960.e20. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.10.019. URL: https://www.sciencedirect.com/science/article/pii/S0092867416314027 (visited on 06/29/2023).

[31] Thomas Panier et al. "Fast Functional Imaging of Multiple Brain Regions in Intact Zebrafish Larvae Using Selective Plane Illumination Microscopy". In: *Frontiers in Neural Circuits* 7 (2013). ISSN: 1662-5110. URL: https://www.frontiersin.org/articles/10.3389/fncir.2013.00065 (visited on 04/07/2023).

[32] Luigi Petrucco et al. "Neural Dynamics and Architecture of the Heading Direction Circuit in Zebrafish". In: *Nature Neuroscience* 26.5 (May 2023), pp. 765–773. ISSN: 1546-1726. DOI: 10.1038/s41593-023-01308-5. URL: https://www.nature.com/articles/s41593-023-01308-5 (visited on 07/27/2025).

[33] Adrián Ponce-Alvarez et al. "Whole-Brain Neuronal Activity Displays Crackling Noise Dynamics". In: *Neuron* 100.6 (Dec. 19, 2018), 1446–1459.e6. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2018.10.045. pmid: 30449656. URL: https://www.cell.com/neuron/abstract/S0896-6273(18)30953-X (visited on 04/05/2023).

[34] Lorenzo Posani et al. "Functional Connectivity Models for Decoding of Spatial Representations from Hippocampal CA1 Recordings". In: *Journal of Computational Neuroscience* 43.1 (Aug. 2017), pp. 17–33. ISSN: 0929-5313, 1573-6873. DOI: 10.1007/s10827-017-0645-9. URL: http://link.springer.com/10.1007/s10827-017-0645-9 (visited on 08/22/2025).

[35] Sebastián A. Romano et al. "Spontaneous Neuronal Network Dynamics Reveal Circuit's Functional Adaptations for Behavior". In: *Neuron* 85.5 (Mar. 4, 2015), pp. 1070–1085. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2015.01.027. URL: https://www.sciencedirect.com/science/article/pii/S0896627315000537 (visited on 04/07/2023).

[36] Xinke Shen et al. "Contrastive Learning of Shared Spatiotemporal EEG Representations across Individuals for Naturalistic Neuroscience". In: *NeuroImage* 301 (Nov. 1, 2024), p. 120890. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2024.120890. URL: https://www.sciencedirect.com/science/article/pii/S1053811924003872 (visited on 07/27/2025).

[37] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. "Learning Compositional Representations of Interacting Systems with Restricted Boltzmann Machines: Comparative Study of Lattice Proteins". In: *Neural Computation* 31.8 (Aug. 1, 2019), pp. 1671–1717. ISSN: 0899-7667. DOI: 10.1162/neco_a_01210. URL: https://doi.org/10.1162/neco_a_01210 (visited on 04/07/2023).

[38] Jérôme Tubiana et al. "Blind Deconvolution for Spike Inference from Fluorescence Recordings". In: *Journal of Neuroscience Methods* 342 (Aug. 1, 2020), p. 108763. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2020.108763. URL: https://www.sciencedirect.com/science/article/pii/S0165027020301862 (visited on 04/07/2023).

[39] Nicholas J. Tustison et al. "The ANTsX Ecosystem for Quantitative Biological and Medical Imaging". In: *Scientific Reports* 11.1 (Apr. 27, 2021), p. 9068. ISSN: 2045-2322. DOI: 10.1038/s41598-021-87564-6. URL: https://www.nature.com/articles/s41598-021-87564-6 (visited on 07/29/2025).

[40] Lucina Q. Uddin, F. Xavier Castellanos, and Vinod Menon. "Resting State Functional Brain Connectivity in Child and Adolescent Psychiatry: Where Are We Now?" In: *Neuropsychopharmacology* 50.1 (Jan. 2025), pp. 196–200. ISSN: 1740-634X. DOI: 10.1038/s41386-024-01888-1. URL: https://www.nature.com/articles/s41386-024-01888-1 (visited on 07/27/2025).

[41] Thijs L van der Plas et al. "Neural Assemblies Uncovered by Generative Modeling Explain Whole-Brain Activity Statistics and Reflect Structural Connectivity". In: *eLife* 12 (Jan. 17, 2023). Ed. by Peter Latham and Laura L Colgin, e83139. ISSN: 2050-084X. DOI: 10.7554/eLife.83139. URL: https://doi.org/10.7554/eLife.83139 (visited on 04/05/2023).

[42] Nikita Vladimirov et al. "Brain-Wide Circuit Interrogation at the Cellular Level Guided by Online Analysis of Neuronal Function". In: *Nature Methods* 15.12 (Dec. 2018), pp. 1117–1125. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0221-x. URL: https://www.nature.com/articles/s41592-018-0221-x (visited on 07/29/2025).

[43] Sébastien Wolf et al. "Sensorimotor Computation Underlying Phototaxis in Zebrafish". In: *Nature Communications* 8.1 (1 Sept. 21, 2017), p. 651. ISSN: 2041-1723. DOI: 10.1038/s41467-017-00310-3. URL: https://www.nature.com/articles/s41467-017-00310-3 (visited on 04/07/2023).

# Supporting information

## Supporting movies

**Movie S1:** Spontaneous whole-brain activity (six fish). Spontaneous, whole-brain, single-neuron binarized spiking activity from six larval zebrafish (Fish 1–Fish 6) recorded with Light-Sheet Microscopy. Shown as a dorsoventral projection, playback is 4× real time.

**Movie S2:** RBM sampling and reconstruction (Fish 1). Sampling of a Restricted Boltzmann machine (RBM) trained on the binarized, whole-brain activity of Fish 1. Left: observed neuronal configuration v. Middle: one RBM Gibbs sampling step $v \to h \to v$. Right: RBM reconstruction $\langle v \rangle$ expected from the observed configuration. Shown in dorsoventral projection.

**Movie S3:** Teacher→Student translation (Fish 3 → Fish 1). Translation of neuronal configurations from the teacher RBM (Fish 3) to the student RBM (Fish 1) through the shared latent space. Left: observed configuration v (teacher, Fish 3). Middle: teacher reconstruction. Right: configuration translated into the student's neuronal space. Dorsoventral projection, playback 4× real time.

**Movie S4:** Student→Teacher translation (Fish 1 → Fish 3). Same as Movie S3, but translating from student to teacher.

All movies are shown in dorsoventral projection. Movies S1, S3, and S4 are accelerated 4×.

## Supporting figures

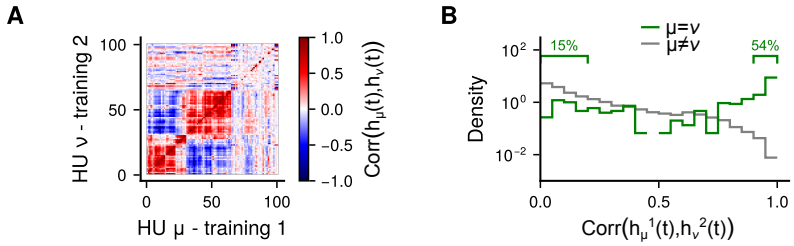



**Fig S1. Supplementary for Fig. 1. A**: Pairwise Pearson Correlation between hidden activity $h_\mu(t) = \mathbb{E}[\mathbf{h} \mid \mathbf{v}_t]$ and $h_\nu(t)$ of two example RBMs trained on the same neuronal recording. The matrix rows have been reordered to obtain the best alignment between the two trainings. **B**: Distribution of correlations from panel A, for best alignments pairs ($\nu = \mu$, green) and other pairs ($\nu \neq \mu$, grey)

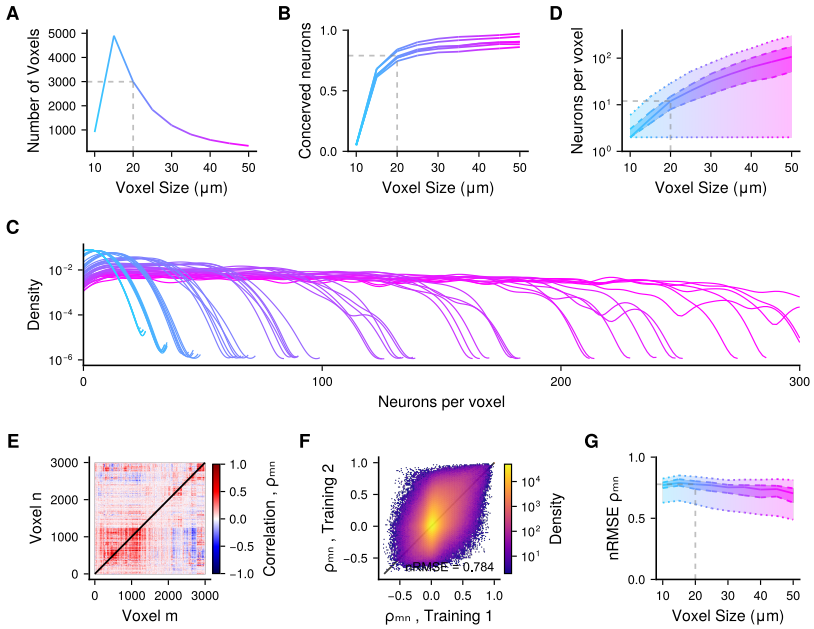



**Fig S2. Supplementary for Fig. 2 investigating the impact of voxel size. A**: Number of populated voxels (each voxel needs to contain $\geq 2$ neurons for each fish, see Methods Voxelised RBMs) as function of voxel size. **B**: Fraction of neurons conserved after voxelization. Each line is a fish. **C**: Distributions of the number of neurons per voxel, for each voxel-size tested (same colors as in panel A), and for each fish (one line per fish). **D**: Median (solid line), 25-75% range (dashed lines), and min-max range (dotted lines) of the number of neurons per voxel (all fish combined). **E**: Pairwise Person correlation coefficients $\rho_{mn}$ between voxels $m$ and $n$ for an example fish at $20\mu m$ voxel size. **F**: Stereotypy in $\rho_{mn}$ between 2 example fish at $20\mu m$ voxel size. **G**: nRMSE (see Methods Measuring identity) of $\rho_{mn}$ between 2 example fish as a function of voxel size.

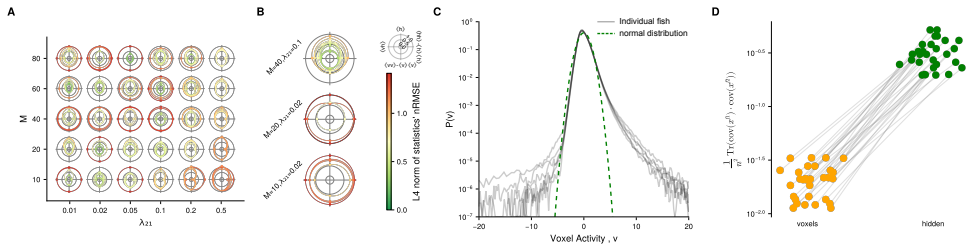



**Fig S3. Supplementary for Fig. 2 cross-validation of voxelized RBM hyperparameters. A**: nRMSE (see Methods Measuring identity) of all four moments (see Methods Evaluation) used to evaluate RBM convergence ($\langle v \rangle$ is omitted as voxelized data is normalized by z-score). Each RBM is represented by a line in polar coordinates where every ray corresponds to a moment $m \in \left\{\langle \mathbf{vh} \rangle, \langle \mathbf{h} \rangle, \langle \mathbf{vv} \rangle - \langle \mathbf{v} \rangle \langle \mathbf{v} \rangle, \langle \mathbf{hh} \rangle - \langle \mathbf{h} \rangle \langle \mathbf{h} \rangle \right\}$, and the ray length corresponds to the $\mathrm{nRMSE}_m(\mathrm{data}, \mathrm{generated})$ between empirical data and data generated by the RBM. Five training were performed and evaluated per hyperparameter pair $(M, \lambda_{21})$, where $M$ is the number of hidden units, and $\lambda_{21}$ is the regularization factor (see Methods Training). All trainings were done on voxelized data from Fish 1. **B**: Same panel A for 3 sets of hyperparameters, and for the concatenated voxelized data of all 6 fish in the dataset. **C**: Distributions of voxelized activity values, with one grey line per fish. Normal distribution in green (dashed). **D**: Fish to fish alignment of eigen vectors in voxelized and hidden space : $\frac{1}{n} \mathrm{Tr}\left(\mathrm{cov}(x^{f1}) \cdot \mathrm{cov}(x^{f2})\right)$, where $x^{f1}$ is activity of one fish, either the voxel activity $v_{tn}$ or the hidden units activity $v_{tn}$, $f1$ and $f2$ are two different fish, and $n$ is either the number of voxels or the number of hidden units.

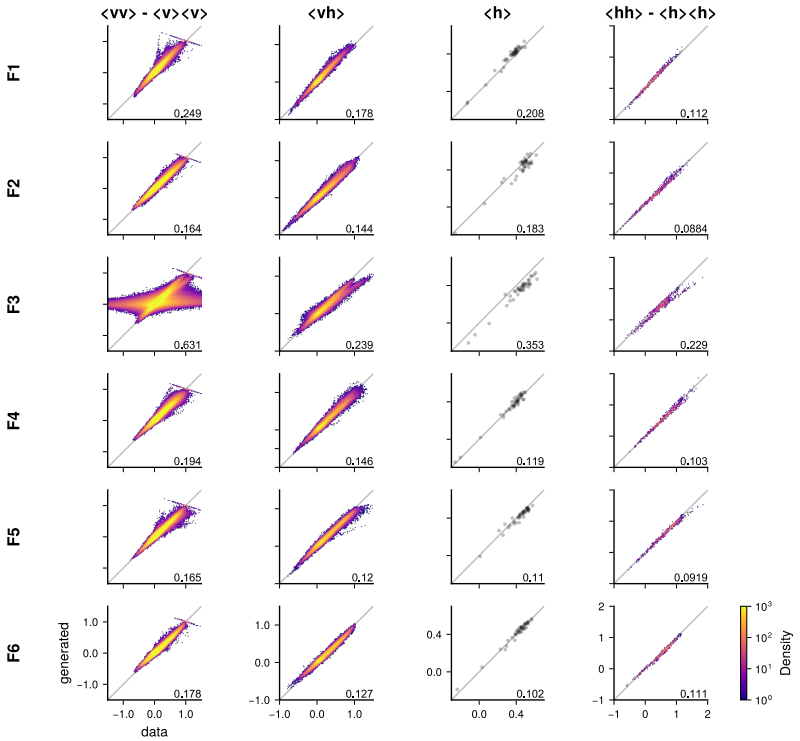



**Fig S4. Supplementary for Fig. 2 showing the voxelized training statistics for each fish.** Identity plots for all 4 moments (see Methods Evaluation) used to evaluate voxelized RBM convergence (notice that moment $\langle v \rangle$ is not represented as voxelized activity is normalized by z-score, see Methods Voxelised RBMs) for each fish.

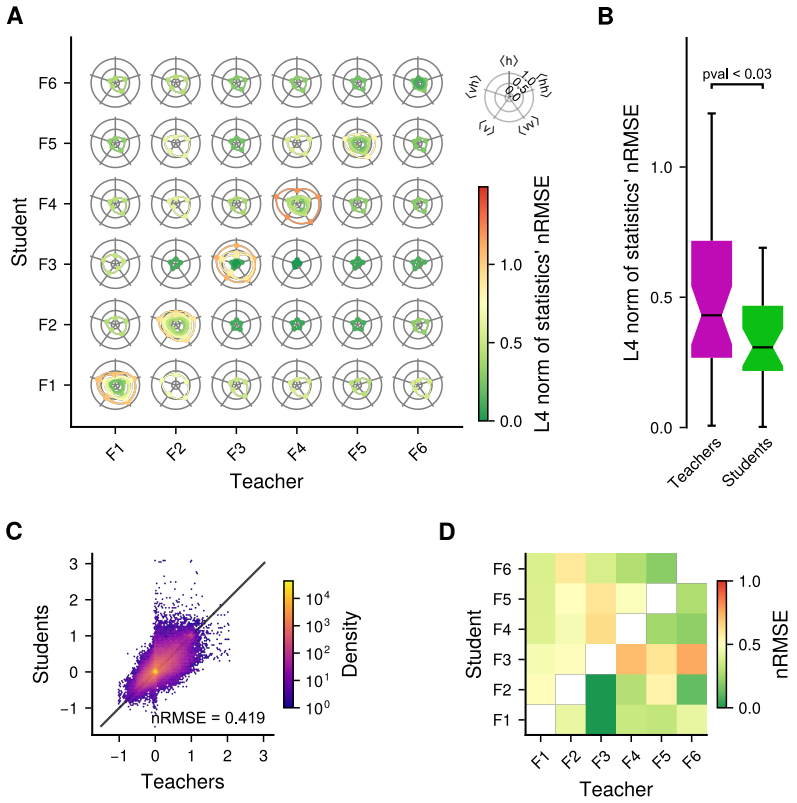



**Fig S5. Supplementary for Fig. 3 showing the training statistics of every teacher-student pairs** **A**: nRMSE (see Methods Measuring identity) of all five moments (see Methods Evaluation) used to evaluate RBM convergence. Each RBM is represented by a line in polar coordinates where every ray corresponds to a moment $m \in \{\langle \mathbf{v} \rangle, \langle \mathbf{vh} \rangle, \langle \mathbf{h} \rangle, \langle \mathbf{vv} \rangle - \langle \mathbf{v} \rangle \langle \mathbf{v} \rangle, \langle \mathbf{hh} \rangle - \langle \mathbf{h} \rangle \langle \mathbf{h} \rangle\}$, and the ray length corresponds to the $\text{nRMSE}_m(\text{data}, \text{generated})$ between empirical data and data generated by the RBM. The line color corresponds to the infinity norm of all five statistics. RBMs along the diagonal correspond to teacher RBMs trained classically (see Methods Standardized RBMs, 10 repetitions). Of-diagonal RBMs are students. (column 3 of this graph corresponds to Fig. S6) **B**: Comparison of training accuracy between teachers (left) and students right, measured as the L4 norm of the moments nRMSE in panel A. The p-value was computed with a one-tailed Mann-Whitney U test. **C**: Identity plot between teacher and student hidden unit pairwise covariance $\langle \mathbf{hh}^\top \rangle - \langle \mathbf{h} \rangle \langle \mathbf{h}^\top \rangle$, for all teacher-student pair combined. **D**: nRMSE of $\langle \mathbf{hh} \rangle - \langle \mathbf{h} \rangle \langle \mathbf{h} \rangle$ for every teacher-student pair.

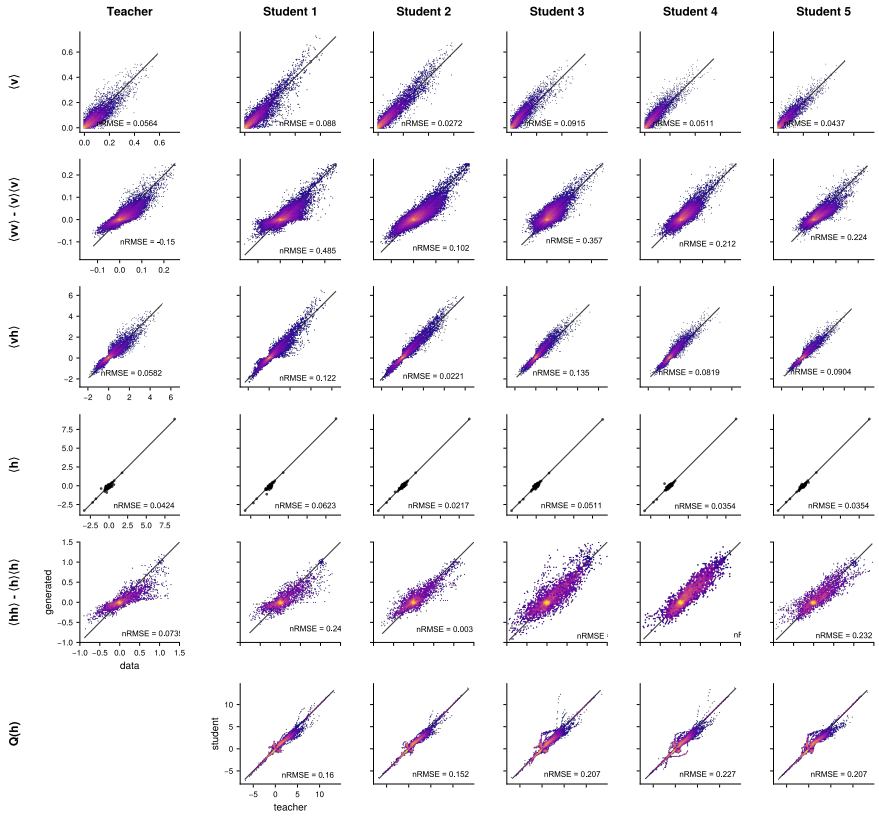



**Fig S6. Supplementary for Fig. 3 showing the training statistics and hidden distribution of all students of the example teacher fish.** Identity plots for all five moments (see Methods Evaluation) used to evaluate RBM convergence and quantile-quantile plot of hidden unit prior distributions $P(h)$ between teacher and students, for the example teacher from Fig. 3 and its five students. (corresponds to column 3 of Fig. S5 A).

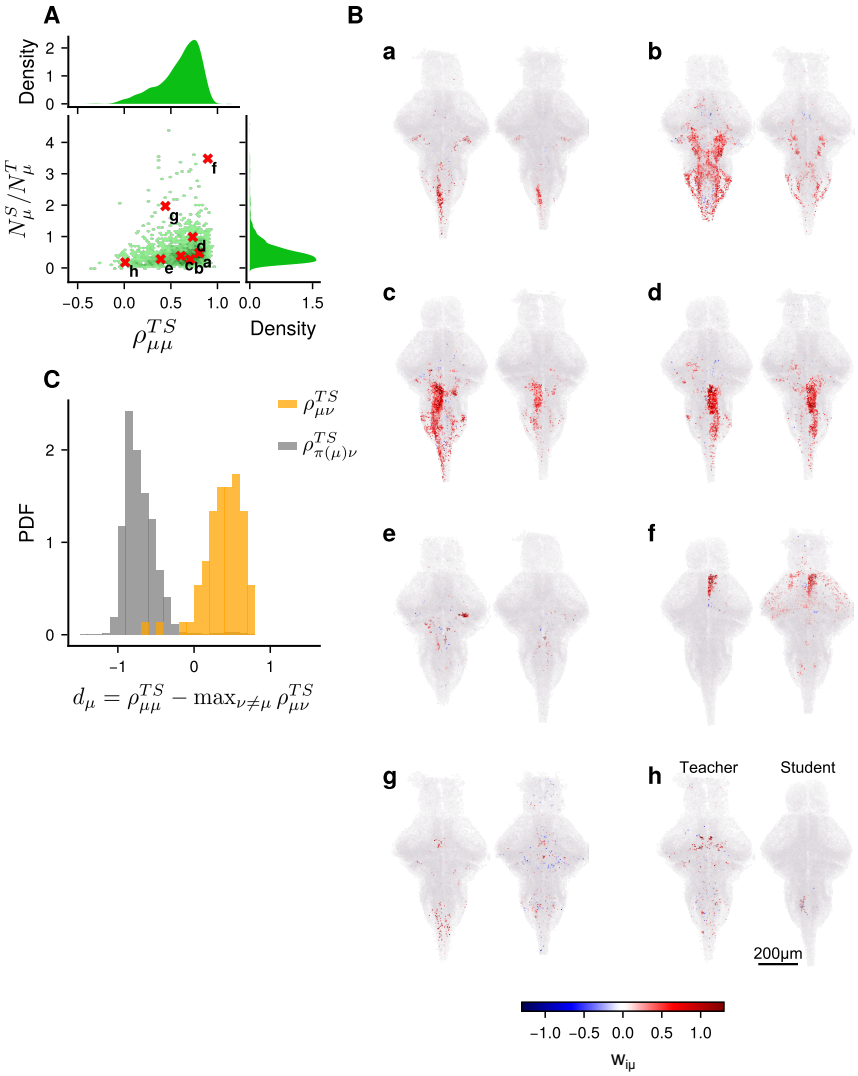



**Fig S7. Supplementary for Fig. 3 on the measurement of distances between weight maps. A**: Joint distribution for the spatial correlation $\rho_{\mu\mu}^{TS}$ (see Methods Measuring Spatial Similarity) of weight maps between teacher and students, and ratio $N_\mu^S/N_\mu^T$ of the number of neurons $N_\mu^T = |w_{i\mu}^T > 10^{-5}|_i$ significantly connected with hidden unit $\mu$ in the teacher and in the student $N_\mu^S$. This distribution was created for all hidden units in all teacher-student pairs. Labeled red crosses correspond to the maps presented in panel B. **B**: Example hidden unit weight maps for teacher (left) and student (right) RBMs. Each map pair corresponds to a red cross in panel A. **C**: Distribution of $d_\mu = \rho_{\mu\mu}^{TS} - \max_{\nu\neq\mu} \rho_{\mu\nu}^{TS}$ (see Methods Measuring Diagonal Dominance) for all hidden units $\mu$ of the example teacher-student pair from Fig. 3. In orange the distribution for the observed matrix $\rho_{\mu\nu}^{TS}$, and in gray for the shuffled matrix $\rho_{\pi(\mu)\nu}^{TS}$.

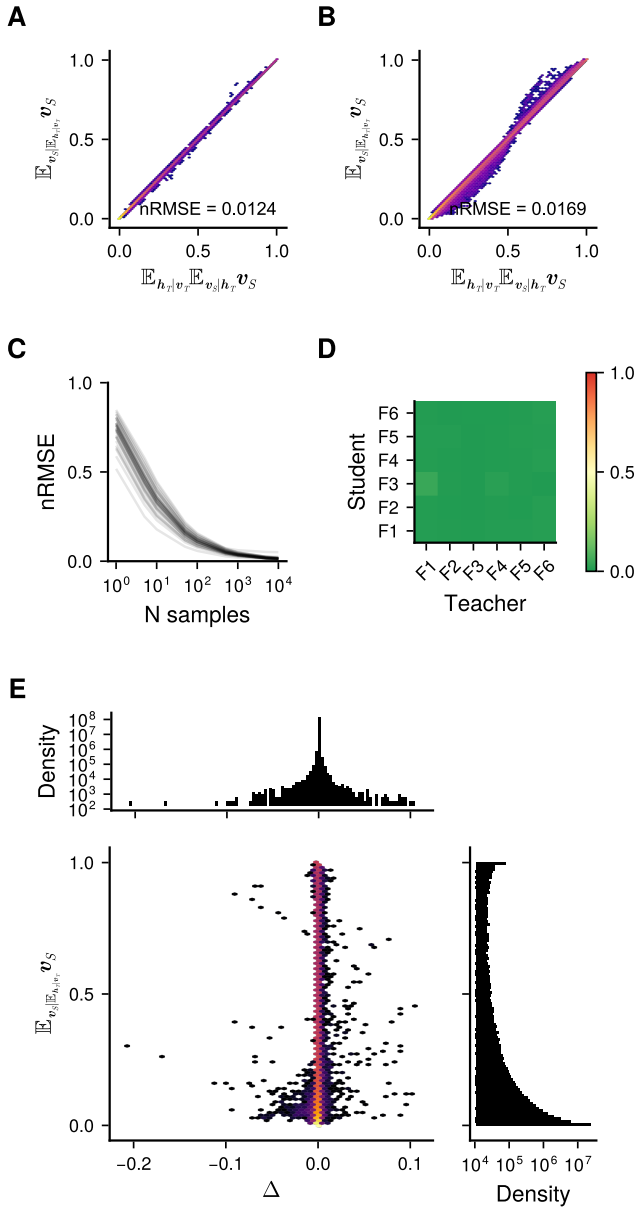



**Fig S8. Supplementary for Fig. 4 and Methods Mapping from one fish to another on the method of translating neuronal configurations between fish.** **A**: Identity plot between the two methods (sampling and expectation) activity transfer presented in Methods Mapping from one fish to another for an example teacher-student RBMs, and 10000 samples. **B**: Same as panel A for all pairs of teacher and student combined. **C**: nRMSE (see Methods Measuring identity) between the 2 methods (see panel A) for all teacher-student pair (one line per pair), as a function of the number of samples used. **D**: nRMSE for activity translation between every teacher-student pair (and activity reconstruction of teacher along the diagonal) for 10000 samples. **E**: Joint distribution of the sampling method and the corrective factor $\Delta$.

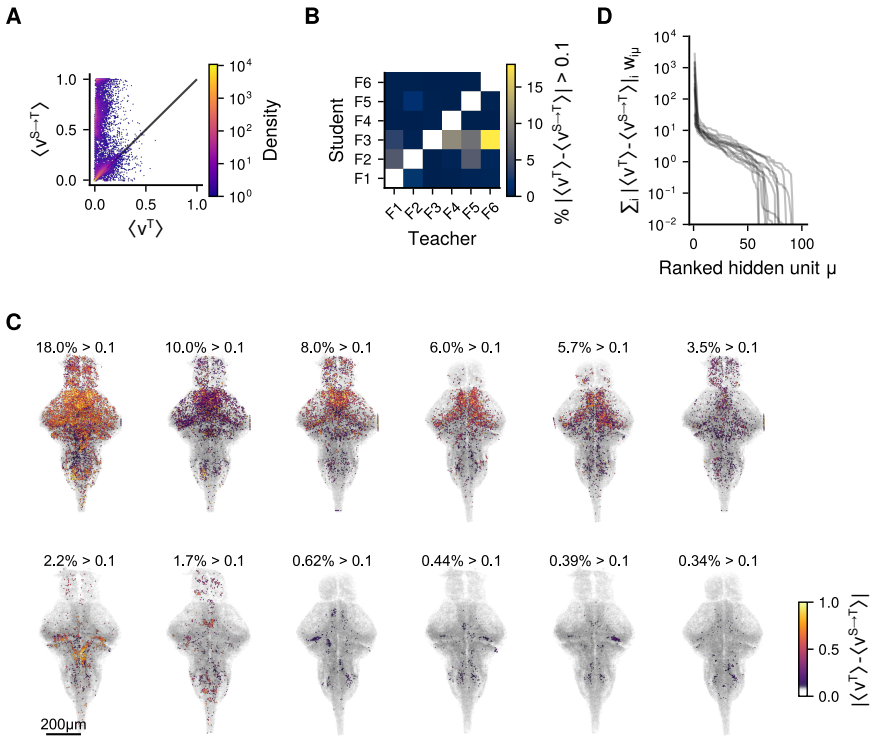



**Fig S9. Supplementary figures for Fig. 4 of the outliers presented in Fig. 4 C,E,G** **A**: Stereotypy of mean neuron activity between an example teacher $\langle \mathbf{v}^T \rangle$ and student translated in to the teacher $\langle \mathbf{v}^{S\to T} \rangle$. We chose specifically the worst outlier. **B**: Percentage of neurons in each student with a translation residual $|\langle \mathbf{v}^T \rangle - \langle \mathbf{v}^{S\to T} \rangle|$ greater that 0.1. **C**: Neuron map of translation residual for the 12 worst teacher-student pairs. **D**: Translation residuals projected onto hidden units $\sum_i |\langle \mathbf{v}^T \rangle - \langle \mathbf{v}^{S\to T} \rangle| w_{i\mu}$. One line per teacher-student pair presented in panel C.

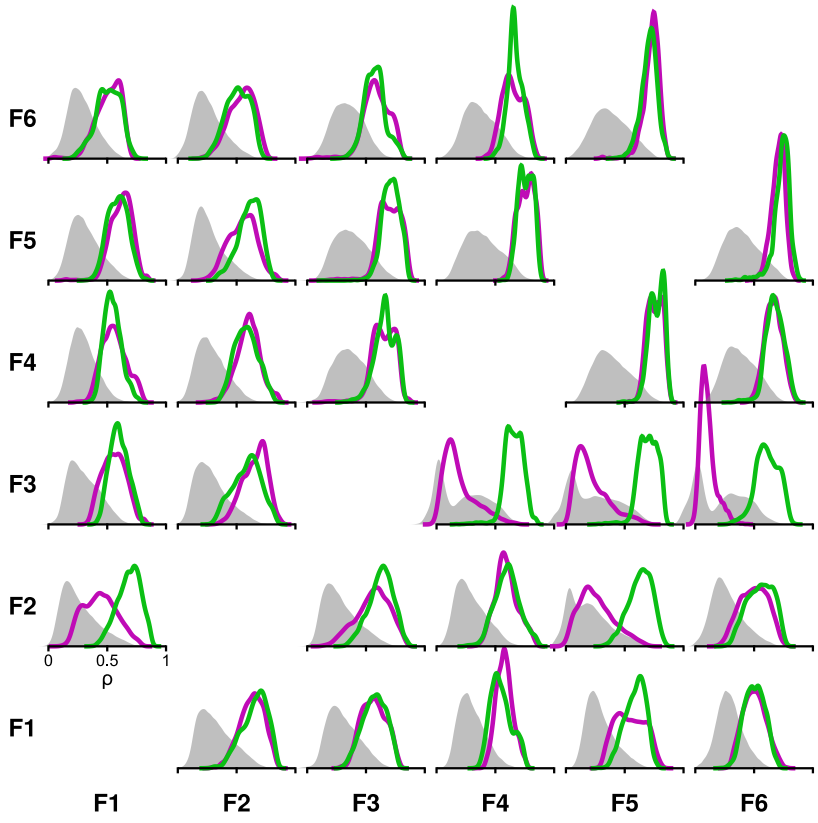



**Fig S10. Supplementary figures for Fig. 4 of the measurement of distances between weight maps.** Same as Fig. 4 F for every teacher-student pair. Spatial correlation $\rho_t^{T\to T,S\to T}$ (magenta) and $\rho_t^{S\to S,T\to S}$ (green) between reconstructed and translated neuronal configuration for each time frame of an example teacher-student pair. In gray, the distribution of spatial correlations between shuffled pairs of time frames.

## Small weight expansion of the LaRBM

Consider a student LaRBM, with the training objective:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\rm v} + \lambda \mathcal{L}_{\rm h} \tag{35}$$

where $\mathcal{L}_{\rm v} = \langle \log P(\mathbf{v}^S_{\rm data})\rangle$ is the log-likelihood of neural activity under the student, $\mathcal{L}_{\rm h} = \langle \log P(\mathbf{h}^T_{\rm sampled})\rangle$ is the log-likelihood of hidden activity sampled from the teacher under the student, and $\lambda \in [0,1]$ is a tunable coefficient. Note that Eq. (35) is simply a rewrite of (2) in the main text.

In this section we consider a second-order expansion of this training objective (35), around small weights. Our goal is to gain some insight on how the trained weights are affected by the combination of the two training objectives, $\mathcal{L}_{\rm v}$ and $\mathcal{L}_{\rm h}$. For simplicity, we will assume that the units have been standardized to have zero means and unit variances.

The expansion of $\mathcal{L}_{\rm v}$ mirrors the expansion of $\mathcal{L}_{\rm h}$, after transposing the weight matrix. Therefore, we will first consider the expansion of $\mathcal{L}_{\rm v}$ in detail. We begin by defining the cumulant generating function of the hidden unit activities,

$$\Gamma_\mu(I_\mu) = \ln \int e^{-\mathcal{U}_\mu(h_\mu) + I_\mu h_\mu} {\rm d}h_\mu \tag{36}$$

where $I_\mu = I_\mu(\mathbf{v}) = \sum_i w_{i\mu} v_i$ are the inputs from the visible to the hidden layer.

Then, the free energy of the visible activity, after marginalizing over the hidden unit configurations, writes:

$$F(\mathbf{v}) = \sum_i \mathcal{V}_i(v_i) - \sum_\mu \Gamma_\mu \left( \sum_i w_{i\mu} v_i \right) \tag{37}$$

The function $\Gamma_\mu(I_\mu)$ is called the cumulant generating function, because the coefficients appearing in its Taylor series are the cumulants of the hidden unit activities at zero weights:

$$\Gamma_\mu(I_\mu) = \kappa^{(0)}_\mu + \kappa^{(1)}_\mu I_\mu + \frac{\kappa^{(2)}_\mu}{2} I_\mu^2 + \dots \tag{38}$$

where $\kappa^{(0)}_\mu = \Gamma_\mu(0) = \int e^{-\mathcal{U}_\mu(h_\mu)} {\rm d}h_\mu$, while $\kappa^{(1)}_\mu, \kappa^{(2)}_\mu, \dots$ are the cumulants of the hidden unit distributions of different orders, at zero weights. For the first few orders, these are the mean, variance, and so on:

$$\kappa^{(1)}_\mu = \frac{\int e^{-\mathcal{U}_\mu(h_\mu)} h_\mu {\rm d}h_\mu}{\int e^{-\mathcal{U}_\mu(h_\mu)} {\rm d}h_\mu}, \quad \kappa^{(2)}_\mu = \frac{\int e^{-\mathcal{U}_\mu(h_\mu)} h_\mu^2 {\rm d}h_\mu}{\int e^{-\mathcal{U}_\mu(h_\mu)} {\rm d}h_\mu} - (\kappa^{(1)}_\mu)^2, \quad \dots \tag{39}$$

It follows that we can expand the free energy $F(\mathbf{v})$ in powers of the weights around the origin as:

$$\begin{aligned} F(\mathbf{v}) &= \sum_i \mathcal{V}_i(v_i) - \sum_\mu \sum_{l=0}^{\infty} \frac{\kappa^{(l)}_\mu}{l!} I^l_\mu(\mathbf{v}) \\ &= \sum_i \mathcal{V}_i(v_i) - \sum_{i\mu} w_{i\mu} v_i \kappa^{(1)}_\mu - \frac{1}{2} \sum_{ij\mu} w_{i\mu} w_{j\mu} v_i v_j \kappa^{(2)}_\mu + \dots \end{aligned} \tag{40}$$

Now let's look at the partition function,

$$\ln Z = \ln \sum_{\mathbf{v},\mathbf{h}} \exp \left( -\sum_i \mathcal{V}_i(v_i) - \sum_\mu \mathcal{U}_\mu(h_\mu) + \sum_{i\mu} w_{i\mu} v_i h_\mu \right) \tag{41}$$

We begin by denoting the measure at zero weights by:

$$P_0(\mathbf{v},\mathbf{h}) = \frac{e^{-\sum_i \mathcal{V}_i(v_i) - \sum_\mu \mathcal{U}_\mu(h_\mu)}}{Z_0}, \quad Z_0 = \sum_{\mathbf{v},\mathbf{h}} e^{-\sum_i \mathcal{V}_i(v_i) - \sum_\mu \mathcal{U}_\mu(h_\mu)} \tag{42}$$

Then we can write:

$$\ln\left(\frac{Z}{Z_0}\right) = \ln\left\langle e^{\sum_{i\mu} w_{i\mu} v_i h_\mu}\right\rangle_0 \tag{43}$$

where $\langle\cdot\rangle_0$ denotes the average under $P_0$. We can treat this expression as a cumulant generating function, which then allows us to write the small weight expansion as follows:

$$\ln(Z) = \ln(Z_0) + \sum_{i\mu} w_{i\mu} \langle v_i h_\mu\rangle_0 + \sum_{ij\mu\nu} \frac{w_{i\mu} w_{j\nu}}{2} \operatorname{cov}_0(v_i h_\mu, v_j h_\nu) + \dots \tag{44}$$

where the averages and covariances are under $P_0$.

We can now substitute the previous equations to obtain the expansion of $\mathcal{L}_\mathrm{v}$ as follows:

$$\begin{aligned}\mathcal{L}_\mathrm{v} = &-\sum_i \langle \mathcal{V}_i(v_i)\rangle^S_\mathrm{data} + \sum_{i\mu} w_{i\mu}\langle v_i\rangle^S_\mathrm{data}\kappa^{(1)}_\mu + \frac{1}{2}\sum_{ij\mu} w_{i\mu} w_{j\mu}\langle v_i v_j\rangle^S_\mathrm{data}\kappa^{(2)}_\mu + \dots \\ &- \ln(Z_0) - \sum_{i\mu} w_{i\mu}\langle v_i h_\mu\rangle_0 - \frac{1}{2}\sum_{ij\mu\nu} w_{i\mu} w_{j\nu} \operatorname{cov}_0(v_i h_\mu, v_j h_\nu) - \dots\end{aligned} \tag{45}$$

We will assume for simplicity that the hidden units are initialized such that $\kappa^{(1)}_\mu = 0$ and $\kappa^{(2)}_\mu = 1$. Then at zero weights it follows that $\langle v_i h_\mu\rangle_0 = 0$. Moreover,

$$\operatorname{cov}(v_i h_\mu, v_j h_\nu) = \langle v_i v_j h_\mu h_\nu\rangle_0 - \langle v_i h_\mu\rangle_0\langle v_j h_\nu\rangle_0 = \langle v_i v_j\rangle_0\langle h_\mu h_\nu\rangle_0 = \delta_{ij}\delta_{\mu\nu} \tag{46}$$

because different units are independent at zero weights. Therefore, omitting terms that are independent of the weights, we arrive at the simplified expansion:

$$\mathcal{L}_\mathrm{v} = \frac{1}{2}\sum_{ij\mu} w_{i\mu} w_{j\mu}\langle v_i v_j\rangle^S_\mathrm{data} - \frac{1}{2}\sum_{i\mu} w^2_{i\mu} - \dots \tag{47}$$

Mirroring this derivation, we obtain a similar expansion for $\mathcal{L}_\mathrm{h}$:

$$\mathcal{L}_\mathrm{h} = \frac{1}{2}\sum_{i\mu\nu} w_{i\mu} w_{i\nu}\langle h_\mu h_\nu\rangle^T_\mathrm{sampled} - \frac{1}{2}\sum_{i\mu} w^2_{i\mu} - \dots \tag{48}$$

where we assume that the visible units are initialized with zero means and unit variances, for simplicity.

Combining the two:

$$\begin{aligned}\mathcal{L} = &(1-\lambda)\left\{\frac{1}{2}\sum_{ij\mu} w_{i\mu} w_{j\mu}\langle v_i v_j\rangle^S_\mathrm{data} - \frac{1}{2}\sum_{i\mu} w^2_{i\mu}\right\} \\ &+ \lambda\left\{\frac{1}{2}\sum_{i\mu\nu} w_{i\mu} w_{i\nu}\langle h_\mu h_\nu\rangle^T_\mathrm{sampled} - \frac{1}{2}\sum_{i\mu} w^2_{i\mu}\right\} \\ = &\frac{1}{2}\sum_{ij\mu\nu} w_{i\mu} w_{j\nu}\left\{(1-\lambda)\langle v_i v_j\rangle^S_\mathrm{data}\delta_{\mu\nu} + \lambda\delta_{ij}\langle h_\mu h_\nu\rangle^T_\mathrm{sampled}\right\} - \frac{1}{2}\sum_{i\mu} w^2_{i\mu}\end{aligned} \tag{49}$$

We now try to understand the eigenmode structure of the matrix $\mathcal{M}_{ij}^{\mu\nu} = (1-\lambda)\langle v_i v_j\rangle_{\text{data}}^S \delta_{\mu\nu} + \lambda\delta_{ij}\langle h_\mu h_\nu\rangle_{\text{sampled}}^T$. If $x_{i\mu}$ is an eigenvector with the eigenvalue $\tau$,

$$\sum_{j\nu} \mathcal{M}_{ij}^{\mu\nu} x_{j\nu} = \tau x_{i\mu} \tag{50}$$

or

$$(1-\lambda)\sum_j \langle v_i v_j\rangle_{\text{data}}^S x_{j\mu} + \lambda \sum_\nu \langle h_\mu h_\nu\rangle_{\text{data}}^T x_{i\nu} = \tau x_{i\mu} \tag{51}$$

We search for eigenvectors of the factorized form $x_{i\mu} = x_i x_\mu$. Substituting:

$$(1-\lambda)\sum_j \langle v_i v_j\rangle_{\text{data}}^S x_j x_\mu + \lambda x_i \sum_\nu \langle h_\mu h_\nu\rangle_{\text{sampled}}^T x_\nu = \tau x_i x_\mu \tag{52}$$

We then assume that $x_i$ is an eigenvector of $\langle v_i v_j\rangle_{\text{data}}^S$, with eigenvalue $\tau_{\text{v}}$, and that $x_\mu$ is an eigenvector of $\langle h_\mu h_\nu\rangle_{\text{sampled}}^T$ with eigenvalue $\tau_{\text{h}}$. Then,

$$(1-\lambda)\tau_{\text{v}} x_i x_\mu + \lambda\tau_{\text{h}} x_i x_\mu = \tau x_i x_\mu$$

or $(1-\lambda)\tau_{\text{v}} + \lambda\tau_{\text{h}} = \tau$. Since $\langle v_i v_j\rangle_{\text{data}}^S$ has $N$ eigenvectors and $\langle h_\mu h_\nu\rangle_{\text{sampled}}^T$ has $M$ eigenvectors, this construction yields all $NM$ eigenvectors of $\mathcal{M}_{ij}^{\mu\nu}$.

This derivation allows us to offer some insight into why the student LaRBM weights often have smaller magnitude than the weights of the teacher RBM. We find empirically that the eigenvalues of $\langle h_\mu h_\nu\rangle_{\text{sampled}}^T$ are often of lower magnitude than the eigenvalues of the covariance matrix of the data used to train the student, $\langle v_i v_j\rangle_{\text{data}}^S$ (see Fig. S11). This implies that the histogram of eigenvalues of the Hessian matrix of the combined objective $\mathcal{L}$ for the student will be shifted to lower values. The corresponding directions in weight space are then more strongly suppressed by the regularization, and will thus be lower.

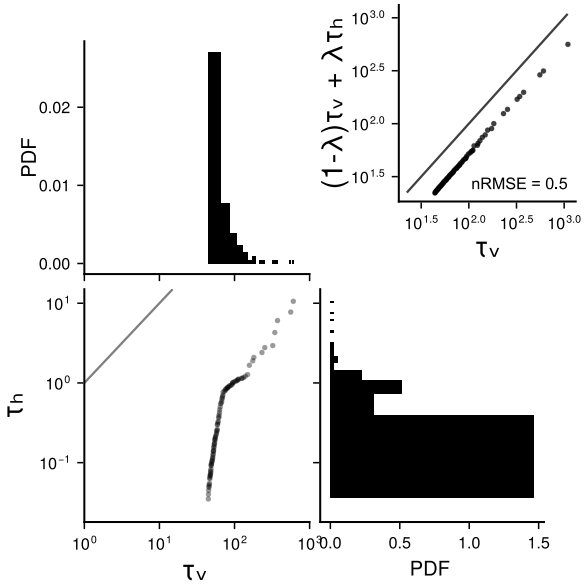



**Fig S11. Supplementary figures for Section Small weight expansion of the LaRBM on the eigenvalues of of the LaRBM training objective.** Top left: histogram of the top 99 eigen values $\tau_{\mathrm{v}}$ of the correlation matrix $\langle v_i v_j \rangle^S_{\mathrm{data}}$ of visible data in the student. Bottom right: histogram of the top 99 eigen values $\tau_{\mathrm{h}}$ of the correlation matrix $\langle h_\mu h_\nu \rangle^T_{\mathrm{sampled}}$ of hidden configuration generated from the teacher RBM. Bottom left: Joint scatter of the $\tau_{\mathrm{v}}$ and $\tau_{\mathrm{h}}$ from the previous two panels. Top Right: QQ plots of the distributions $\tau_{\mathrm{v}}$ and $(1-\lambda)\tau_{\mathrm{v}} + \lambda\tau_{\mathrm{h}}$ with $\lambda = 0.5$.