LEARNING MULTIMODAL POPULATION RELATIONS FOR CROSS-RECORDING NEURON MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Establishing reliable neuronal correspondences across recordings is essential for comparing neural dynamics across individuals and linking functional activity to circuit organization. However, this remains challenging because neuronal identities cannot be reliably determined from individual features alone, as spatial organization and functional activity patterns vary across recordings. Here, we propose NeuReL, a supervised neuron matching framework that integrates multimodal population relations by jointly learning positional and functional relationships among neurons within their surrounding populations. NeuReL first captures population-level representations within each recording and then refines correspondences by enforcing relational consistency across recordings. On held-out animals, NeuReL achieves significant improvements over the strongest evaluated baselines across multiple *C. elegans* datasets. We further validate NeuReL on longitudinal neuron matching in zebrafish, demonstrating its applicability across species and recording conditions. **These results demonstrate that multimodal population relations provide a powerful basis for cross-recording neuron matching**, enabling reliable neuronal correspondences for comparing neural activity across sessions and individuals.

1 INTRODUCTION

Neuron matching across recordings is essential for understanding how neural dynamics relate to circuit organization. Large-scale calcium imaging now enables simultaneous recording of activity from many individual neurons (Schrödel et al., 2013; Nguyen et al., 2016; Venkatachalam et al., 2016), creating opportunities to compare neural dynamics across recordings. However, reliable matching remains challenging even in nervous systems with well-defined neuronal identities. The adult hermaphrodite *C. elegans* has 302 neurons with canonical names (White et al., 1986; Yemini et al., 2021), but their spatial positions vary with posture, nonrigid deformation, and individual differences, and are further affected by imaging and segmentation errors (Yu et al., 2021; Wu et al., 2022). Neural activity provides additional information for distinguishing neurons, but also varies with behavioral state, population state, and recording conditions (Ji et al., 2021). Moreover, different recordings may contain only partially overlapping neuronal populations.

Existing neuron-matching approaches have primarily relied on spatial geometry (Yu et al., 2021; Wu et al., 2022) or activity-derived information (Lanza et al., 2024; Kamiya et al., 2025), sometimes exploiting population structure within a single modality. Recent representation-learning approaches have further shown that information from surrounding neuronal populations can improve the inference of neuronal attributes such as cell type and brain region (Mi et al., 2023; Arora et al., 2025). However, these attributes can be shared by multiple neurons and therefore do not establish individual correspondences across recordings. This motivates a population-relative view of neuron correspondence: even when individual-neuron features are ambiguous, complementary geometric and functional relationships to the surrounding population may provide additional evidence for distinguishing corresponding neurons.

To address this problem, we propose **Neural Relational Learning (NeuReL)**, a supervised framework for matching individual neurons across recordings. Using three-dimensional neuronal positions and neural activity traces, NeuReL jointly learns representations of individual neurons and their relations to the surrounding population. These relations are incorporated into neuron represen-

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

tations and used to refine cross-recording matches based on population-level relational consistency. Since activity relations are computed within each recording, NeuReL does not require temporal alignment across recordings. We demonstrate the utility of population relations for neuron matching and further show that neural activity provides identity information complementary to position, particularly for distinguishing neurons with similar spatial features. Our main contributions are:

1. We introduce a new perspective for cross-recording neuron matching by **representing neurons through multimodal population relations**, integrating geometric and functional information beyond individual-neuron features.
2. We develop a **differentiable optimal transport-based matching framework** that optimizes cross-recording correspondences through population-level relational consistency while handling unmatched neurons.
3. We demonstrate the **robustness of NeuReL across experimental conditions and species**, achieving reliable cross-animal matching in *C. elegans* and longitudinal neuron matching in zebrafish.
4. Our results support a general principle for neuron correspondence: **combining complementary modalities with relationships to surrounding neurons** provides more reliable matching evidence than either source alone.

2 Related Work

Neuron matching. Neuron matching has traditionally relied primarily on spatial information. Classical approaches align neurons across recordings through point-set registration; for example, CPD uses non-rigid transformations to account for geometric deformation (Myronenko & Song, 2010). For cross-animal matching, variability in neuronal organization across individuals has further motivated the use of population-level information to constrain identity assignments. StatAtlas learns distributions of neuronal positions and fluorescent colors from annotated animals (Varol et al., 2020), CRF_ID uses relative spatial relationships between neurons to constrain identity assignments (Chaudhary et al., 2021), and fDNC uses a Transformer to learn population-level spatial context from neuronal coordinates (Yu et al., 2021). In general point-set matching, GeoTransformer (Qin et al., 2022) likewise exploits geometric relations for correspondence estimation. These approaches demonstrate the value of population-level spatial structure beyond individual neuronal positions. However, spatial information can remain ambiguous when multiple candidates provide similar geometric evidence.

Neural activity provides complementary identity information when spatial evidence is ambiguous. Previous work has combined neuronal position with predefined activity patterns to identify a subset of *C. elegans* neurons with stereotyped activity (Lanza et al., 2024). Other work derives relations from spontaneous population activity and performs unsupervised matching using Gromov–Wasserstein alignment (Kamiya et al., 2025). NeuReL jointly learns geometric and activity relations under correspondence supervision, using both modalities to establish individual neuron correspondences across recordings.

Neuron representations. Population context is also useful for learning representations of individual neurons from neural activity. NeuPRINT uses statistics of population activity to learn neuronal representations that are relatively invariant over time (Mi et al., 2023), NeurPIR represents neurons using learned population features (Wu et al., 2025), and NuCLR uses self-supervised spatiotemporal modeling to integrate activity across neurons (Arora et al., 2025). These studies demonstrate how the surrounding population can inform neuronal representations and support the prediction of attributes such as cell type or brain region. However, multiple neurons can share these attributes, so predicting them does not establish individual correspondences across recordings. NeuReL extends this population-context perspective to individual neuron matching by learning contextualized neuron representations while retaining explicit pairwise relations for subsequent matching.

Relational matching. Using population relations to evaluate correspondences connects neuron matching to the broader problem of relational set matching. Standard optimal transport estimates correspondences from costs between individual elements (Cuturi, 2013), whereas Gromov–Wasserstein compares pairwise relations within two sets to perform structural matching (Peyré

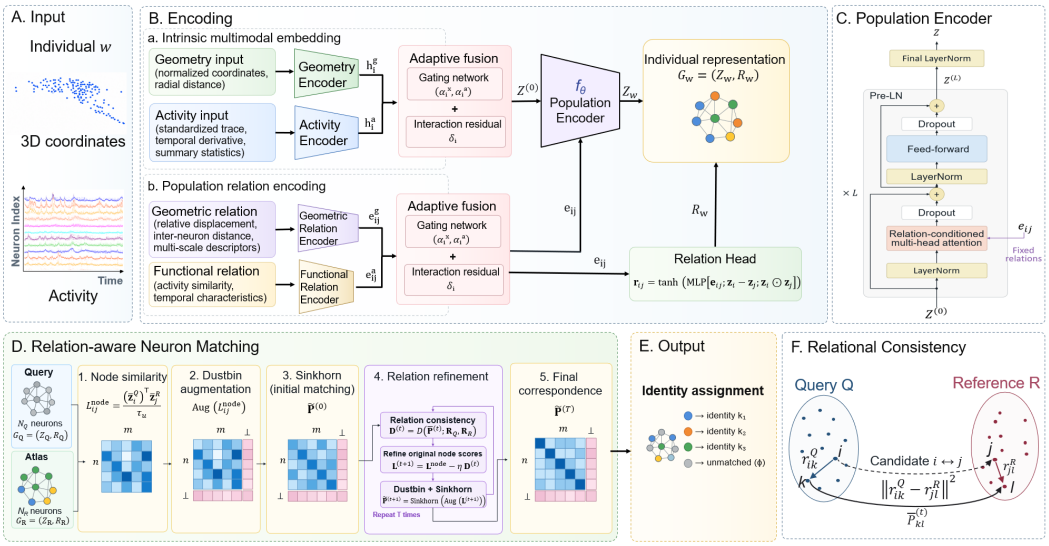

Figure 1: **Overview of the NeuReL framework. (A)** Each recording provides 3D neuronal coordinates and activity traces. **(B,C)** Geometry and activity embeddings are adaptively fused at both the node and relation levels. A relation-conditioned Transformer contextualizes node representations, while a relation head produces pairwise relation vectors. **(D,E)** Query neurons are matched either to another recording or to a population atlas constructed from training recordings and fixed at inference. Node similarity initializes a transport plan through Sinkhorn normalization with explicit dustbin states. Iterative refinement based on relational consistency produces identity assignments or unmatched predictions. **(F)** For a candidate match $i \leftrightarrow j$, the discrepancy between $\mathbf{r}^Q_{ik}$ and $\mathbf{r}^R_{jl}$ is weighted by the current soft correspondence $\bar{P}^{(t)}_{kl}$ supporting $k \leftrightarrow l$.

et al., 2016). Fused Gromov–Wasserstein (FGW) combines individual features with relational structure (Vayer et al., 2020), while its unbalanced extension, FUGW, accommodates partially overlapping sets (Thual et al., 2022). These formulations provide a basis for matching with feature and relational information. NeuReL combines supervised relation learning with iterative correspondence refinement. It updates candidate matches by directly comparing relations across recordings and incorporates an explicit unmatched state for neurons without reference counterparts.

3 Method

The goal of this work is to identify corresponding neurons across individuals and across repeated recordings of the same individual using neuronal geometry and activity. Our approach is grounded in the idea that a neuron's identity is reflected not only in its own geometry and activity, but also in its positional and functional relations to the population. These relations can provide relatively comparable evidence for correspondence even when individual features vary across recordings. Based on this idea, NeuReL first constructs neuron representations that incorporate population context, then evaluates candidate correspondences using population-level relational consistency. Figure 1 provides an overview of the framework.

Each recording contains a population of observed neurons, each with three-dimensional coordinates and an activity trace. We match a query recording to a reference consisting of either observed neurons from another recording or identity prototypes in an atlas. To accommodate neurons without a counterpart in the reference, each query neuron can be matched to a real reference candidate or assigned to an unmatched state $\perp$.

Multimodal population relation encoding. We aim to retain explicit relations for subsequent matching. We begin by encoding each neuron's geometry and activity separately. The activity encoder processes the standardized activity trace and its first temporal difference as two input channels

using a temporal CNN, and combines the resulting features with summary statistics computed before standardization. This allows the encoder to capture both temporal dynamics and local changes while preserving information about overall activity level and variability. Because the relative reliability of geometry and activity can vary across neurons, a gating MLP predicts neuron-specific modality scores that are converted by a softmax into geometry and activity weights. The weighted embeddings are further augmented with an interaction residual that captures information arising from their joint representation. Their fusion produces the initial node representations.

To describe the spatial and functional relations of neurons relative to the population, we construct geometric and activity relation descriptors for each ordered neuron pair (i, j). The geometric relations are used to preserve directional information and characterize neuronal proximity at different scales. The activity relations are used to describe similarity in overall activity patterns and in temporal changes. The geometric and activity relations are transformed by separate encoders and then adaptively fused into a unified relation representation.

We use a relation-conditioned Transformer (Vaswani et al., 2017; Shaw et al., 2018) to contextualize each neuron's initial representation by aggregating information from the surrounding population. For the input $\mathbf{h}_i$ to a layer, let $\widehat{\mathbf{h}}_i = \mathrm{LN}(\mathbf{h}_i)$. Attention head h computes

$$
\begin{aligned}
s_{ij}^h &= \frac{Q_h(\widehat{\mathbf{h}}_i)^\top K_h(\widehat{\mathbf{h}}_j)}{\sqrt{d_h}} + b_h(\mathbf{e}_{ij}), \\
\mathbf{u}_i^h &= \sum_j \mathrm{softmax}_j(s_{ij}^h) \left(V_h(\widehat{\mathbf{h}}_j) + \psi_h(\mathbf{e}_{ij}) \right),
\end{aligned}
\tag{1}
$$

where Q_h, K_h, and V_h are learned affine projections and d_h is the head dimension. The learned linear maps b_h and ψ_h make relations influence both how strongly neuron i attends to neuron j and what information it receives from j. This yields the contextualized embeddings $\mathbf{z}_i$.

Finally, comparing node embeddings alone does not directly compare how neurons relate to others in their respective recordings. We therefore construct a directed relation vector $\mathbf{r}_{ij}$ that combines the pairwise relation embedding with the population-context information received by the two neurons. Each recording w is represented by node embeddings $\mathbf{Z}_w = \mathbf{z}_i$ and relation vectors $\mathbf{R}_w = \mathbf{r}_{ij}$. The matching algorithm uses the node embeddings to initialize matches and compares the relation vectors to refine them.

Relation-aware neuron matching. The two populations may contain different subsets of neurons, so we allow neurons without a counterpart to remain unmatched. Given a query $(\mathbf{Z}_Q, \mathbf{R}_Q)$ with n neurons and a reference $(\mathbf{Z}_R, \mathbf{R}_R)$ with m candidates, we use a transport plan $\widetilde{\mathbf{P}} \in \mathbb{R}_+^{(n+1)\times(m+1)}$ whose entries are soft matching weights. The extra row and column represent an unmatched state $\perp$ on each side. We initialize these weights by comparing the contextualized neuron embeddings:

$$
\begin{aligned}
L_{ij}^{\mathrm{node}} &= \frac{(\bar{\mathbf{z}}_i^Q)^\top \bar{\mathbf{z}}_j^R}{\tau_u}, \\
\widetilde{\mathbf{P}}^{(0)} &= \mathrm{Sinkhorn}\big(\mathrm{Aug}(\mathbf{L}^{\mathrm{node}}); \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\nu}}\big).
\end{aligned}
\tag{2}
$$

Here $\bar{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$ is the ℓ_2-normalized node embedding, making the dot product a cosine similarity. This removes the influence of embedding magnitude on the initial matching scores, while the learned temperature $\tau_u > 0$ controls their scale. The operator Aug appends the unmatched row and column, each with a learned scalar score. Sinkhorn normalization (Cuturi, 2013) adjusts the weights using target row sums $\widetilde{\boldsymbol{\mu}}$ and column sums $\widetilde{\boldsymbol{\nu}}$. These targets give equal total weight to each real neuron and allow each unmatched state to receive weight from multiple neurons.

Corresponding neurons are expected to preserve consistent population relations with their surrounding neurons across recordings. Therefore, after obtaining an initial match based on node representations, we directly compare relational structures across recordings and iteratively refine candidate correspondences according to their relational consistency. Specifically, for a candidate match $i \leftrightarrow j$, we compare the relation from query neuron i to k with that from reference candidate j to l, as illustrated in Figure 1. Pairs $k \leftrightarrow l$ with larger current matching weights contribute more to this comparison. Let $\bar{\mathbf{P}}^{(t)}$ contain the entries of $\widetilde{\mathbf{P}}^{(t)}$ excluding the unmatched row and column, rescaled

to sum to one. We update the matching weights as

$$D_{ij}^{(t)} = \sum_{k,l} \bar{P}_{kl}^{(t)} \left\| \mathbf{r}_{ik}^{Q} - \mathbf{r}_{jl}^{R} \right\|_2^2,$$
$$\widetilde{\mathbf{P}}^{(t+1)} = \mathrm{Sinkhorn}\Big(\mathrm{Aug}\Big(\mathbf{L}^{\mathrm{node}} - \eta \mathbf{D}^{(t)} \Big) ; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\nu}} \Big), \tag{3}$$

where $\eta > 0$ is a learned weight controlling the strength of the relation penalty. The quantity $Dij^{(t)}$ measures the average relation disagreement for matching i to j, weighted by the current assignment. At each refinement step, the relation discrepancy computed from the current soft assignment is used to construct the next matching scores, and Sinkhorn approximately solves the resulting optimal-transport-based soft matching problem. We unroll two such refinement steps rather than iterating to convergence. All refinement and Sinkhorn operations remain differentiable, allowing the final matching loss to backpropagate through the entire procedure.

Training and inference. We train all components on pairs of recordings. To learn when a neuron should remain unmatched, we first establish known correspondences and then randomly remove neurons independently from both recordings. If both neurons in a known pair remain, we supervise their match; if only one remains, we supervise it as unmatched. NeuReL supports partial matching through explicit unmatched states, while random neuron removal during training encourages robustness to missing neurons. All retained neurons, including unlabeled ones, provide population context. Missing labels alone do not create unmatched targets because a counterpart may still be present.

We supervise the matching algorithm in both directions using a focal loss (Lin et al., 2017), which reduces the contribution of neurons whose correct match or unmatched target already receives a high score. Let θ denote all trainable parameters and ξ specify the neurons randomly removed from a training pair (Q, R). The sets $\mathcal{I}_Q$ and $\mathcal{I}_R$ contain the supervised neurons remaining in the query and reference recordings. We optimize

$$\theta^{\star} = \arg\min_{\theta} \; \mathbb{E}_{(Q,R),\,\xi} \left[- \frac{\sum_{i \in \mathcal{I}_Q} (1 - p_{\theta,i}^{Q})^{\gamma} \log p_{\theta,i}^{Q} + \sum_{j \in \mathcal{I}_R} (1 - p_{\theta,j}^{R})^{\gamma} \log p_{\theta,j}^{R}}{|\mathcal{I}_Q| + |\mathcal{I}_R|} \right]. \tag{4}$$

Here the expectation averages over training pairs and random neuron removal. The scores $p_{\theta,i}^{Q}$ and $p_{\theta,j}^{R}$ are assigned to each supervised neuron's known counterpart or unmatched target. We obtain them by dividing the corresponding final matching weight by the target row sum on the query side or target column sum on the reference side, followed by clipping for numerical stability. The exponent γ controls the reduction in weight for targets with high scores. Gradients pass through all refinement and Sinkhorn updates.

We select the checkpoint with the highest bidirectional Top-1 accuracy on validation recording pairs. We then freeze its parameters. To combine identity information across training recordings, we construct a relational atlas by averaging normalized node embeddings for each identity and relation vectors for each jointly annotated identity pair. Identity pairs never jointly annotated receive zero relation vectors, which remain included in the relation penalty. Each query recording is encoded independently and matched to either another recording or this fixed atlas. Test identity labels are used only for evaluation. Additional implementation details are provided in Appendix C.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate NeuReL across distinct imaging conditions using two *C. elegans* datasets for cross-animal neuron matching and one zebrafish dataset for longitudinal matching. For *C. elegans*, we use recordings from 38 freely moving animals in the Atanas dataset (Atanas et al., 2023) and 95 immobilized animals in the Kato dataset (Dunn et al., 2025). Both datasets provide 3D neuronal positions, calcium activity, and neuron identity annotations, with sparser annotations in Kato. We further use longitudinal light-sheet calcium imaging recordings from eight agarose-embedded zebrafish embryos (Wan et al., 2019; Keller, 2019). The functional recordings cover 9–10 spinal

segments during early circuit development rather than the whole brain, and provide calcium activity, 3D spatial trajectories, and tracking-based ground-truth correspondences across developmental time points. Since tracking identities are defined within each fish rather than shared across fish, we match recordings from different time points within each held-out fish. Detailed dataset statistics and experimental protocols are provided in Appendix D (Table 13).

Evaluation Protocol. For both *C. elegans* datasets, we use five-fold cross-validation, keeping all recordings from the same animal in the same fold. Baselines that require a single training recording use the training animal with the smallest average geometric distance to the remaining training animals.

For zebrafish, we use eight-fold leave-one-fish-out cross-validation, with six fish for training, one for validation, and one for testing in each fold. We evaluate longitudinal matching between pairs of recordings from the held-out fish, using cell tracking to establish ground-truth correspondences. Each test pair comprises two 30-second recording windows whose start times are separated by 60 minutes; we evaluate 101 such pairs across the eight folds. For StatAtlas and CRF_ID, reference models are constructed from recordings at other time points of the same held-out fish using their tracking identities, excluding both recordings in each test pair.

For each test–reference pair r, we compute accuracy as

$$\mathrm{Acc}_r = \frac{\#\text{correct ground-truth matches in pair } r}{\#\text{ground-truth matches in pair } r}. \tag{5}$$

Top-1 accuracy uses the highest-scoring reference candidate independently for each test neuron, whereas Hungarian accuracy uses the maximum-score one-to-one assignment. We report the unweighted mean across all test–reference pairs. Mean Reciprocal Rank (MRR) is computed analogously for each pair and then averaged across pairs.

Baselines. Given the novelty of our problem setting, there are no "fair" baselines that directly match our use of both neuronal geometry and neural activity for cross-recording neuron matching under the same evaluation setting. We therefore compare NeuReL with representative baselines spanning Transformer-based matching (fDNC and GeoTransformer), graph-based matching (CRF_ID), activity-based representation learning (NuCLR), optimal transport (FUGW), and classical registration or statistical atlases (CPD and StatAtlas). We adapt FUGW to incorporate both modalities, using geometry to define cross-population matching costs and neural-activity distances to characterize within-population structure. Implementation details and hyperparameter settings are provided in Appendix E.

4.2 Neuron Matching Across Recordings

***C. elegans*.** As shown in Table 1, NeuReL achieves the highest mean Top-1 and Hungarian accuracies on both datasets. On Atanas, NeuReL reaches 78.06% Top-1 accuracy, exceeding the strongest baseline, CPD (48.34%), by 29.72 percentage points. On Kato, it reaches 74.62%, outperforming the strongest baseline, StatAtlas (48.39%), by 26.23 points. NeuReL also achieves Hungarian accuracies of 78.27% on Atanas and 74.67% on Kato, improving over the corresponding strongest baselines, CPD (51.70%) and StatAtlas (50.49%), by 26.57 and 24.18 percentage points, respectively. These gains show that NeuReL's advantage holds under both independent prediction and one-to-one assignment.

Zebrafish. As shown in Table 2, NeuReL achieves the highest mean Top-1 accuracy (97.35%), MRR (0.9844), and Hungarian accuracy (97.28%). Compared with CRF_ID, the strongest baseline across all three metrics, NeuReL improves Top-1 and Hungarian accuracies by 0.94 and 0.69 percentage points, respectively, while increasing MRR from 0.9805 to 0.9844. Geometry-only methods already perform strongly in this longitudinal matching setting. In contrast, activity alone may not provide sufficiently distinctive information for neuron matching. Our within-recording analysis shows that approximately 24.7% of neurons have at least one other neuron with a highly correlated activity trace (Pearson $r > 0.8$). This similarity between distinct neurons may lead to ambiguity for activity-only matching and contribute to NuCLR's weaker performance. Changes in activity across recording times or neural states could further reduce the cross-recording stability of activity-based representations.

Table 1: Cross-animal neuron matching on held-out *C. elegans* recordings. Values are the unweighted mean ± sample SD across test–reference pairs (%). Atanas includes 38 pairs; Kato includes 87 valid pairs with at least one ground-truth match. Best results in each column are shown in bold.

Type	Method	Atanas		Kato	
		Top-1 ↑	Hung. ↑	Top-1 ↑	Hung. ↑
Geometry only	CPD	48.34 ± 14.13	51.70 ± 14.52	22.12 ± 31.10	25.11 ± 34.25
	fDNC	42.74 ± 17.10	46.85 ± 17.57	40.31 ± 28.14	45.66 ± 33.15
	GeoTransformer	46.92 ± 13.22	42.99 ± 13.20	33.89 ± 27.30	19.20 ± 19.83
	StatAtlas	12.44 ± 6.32	12.08 ± 7.29	48.39 ± 33.08	50.49 ± 35.86
	CRF_ID	24.61 ± 9.85	24.61 ± 9.85	5.32 ± 13.63	5.32 ± 13.63
Activity only	NuCLR	21.81 ± 5.94	22.53 ± 6.37	11.09 ± 18.59	12.80 ± 16.37
Geometry + Activity	FUGW	32.11 ± 12.62	37.00 ± 13.64	23.14 ± 26.98	8.60 ± 16.91
	NeuReL (Ours)	**78.06 ± 16.92**	**78.27 ± 17.08**	**74.62 ± 26.85**	**74.67 ± 26.83**

Table 2: Longitudinal neuron matching on held-out zebrafish. Values are the unweighted mean ± sample SD across 101 test–reference pairs. Top-1 and Hungarian accuracies are reported as percentages, while MRR is reported on a 0–1 scale. Best results in each column are shown in bold.

Type	Method	Top-1 ↑	MRR ↑	Hung. ↑
Geometry only	CPD	89.64 ± 6.86	0.9417 ± 0.0418	93.96 ± 6.62
	fDNC	82.90 ± 14.44	0.9042 ± 0.0914	94.98 ± 10.63
	GeoTransformer	69.95 ± 12.19	0.7494 ± 0.1182	70.84 ± 14.77
	StatAtlas	87.54 ± 24.51	0.9083 ± 0.2170	88.70 ± 25.01
	CRF_ID	96.41 ± 4.13	0.9805 ± 0.0233	96.59 ± 3.96
Activity only	NuCLR	6.59 ± 3.25	0.1766 ± 0.0458	6.60 ± 3.86
Geometry + Activity	FUGW	68.33 ± 10.89	0.8094 ± 0.0727	82.51 ± 10.91
	NeuReL (Ours)	**97.35 ± 6.95**	**0.9844 ± 0.0448**	**97.28 ± 7.21**

Across all three datasets, NeuReL consistently achieves the highest mean Top-1 and Hungarian accuracies, whereas the strongest baseline varies across datasets. Geometry-only methods perform well in longitudinal zebrafish matching but achieve lower accuracies in cross-animal *C. elegans* matching. One possible explanation is that spatial variability across animals makes geometry alone less reliable for resolving neuronal identity. NeuReL's larger gains on the two *C. elegans* datasets suggest that its multimodal relational formulation is particularly effective in this cross-animal setting, while remaining competitive when geometry already provides strong matching evidence.

4.3 Activity Helps Distinguish Spatially Ambiguous Neurons

We next ask whether activity is most helpful for neurons that are difficult to distinguish geometrically on Atanas. For each query, we compare how closely its activity relationships to surrounding neurons match those of the correct identity versus three spatially nearby incorrect candidates (Hard candidates). We average this difference across animal pairs for each identity. Identities for which activity more clearly separates the correct neuron from nearby incorrect candidates tend to show larger Top-1 gains when activity is added to NeuReL (Figure 2(A); Spearman $\rho = 0.247$, $p = 0.002$).

Six selected cases illustrate the complementary information provided by activity relations (Figure 2(B)). In each case, the model without activity predicts a spatial Hard candidate, whereas NeuReL identifies the correct neuron. The correct identity and the nearby incorrect candidates have similarly matched spatial relationships to surrounding neurons. In contrast, the correct identity shows the most similar activity relationships to the surrounding population. Activity therefore provides additional information for distinguishing neurons that are difficult to separate using geometry alone. The heatmaps show these similarities relative to a single training recording, whereas the final model predictions are made using the population atlas.

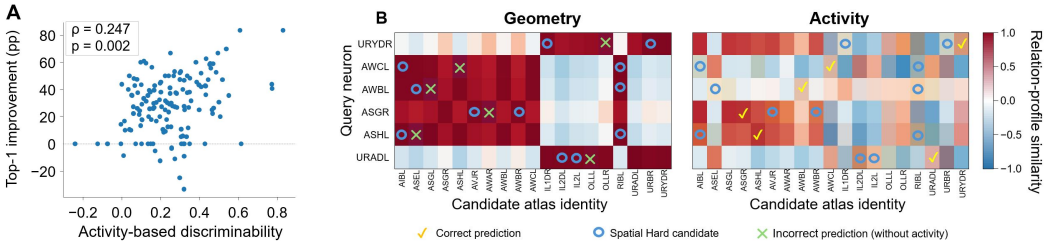

Figure 2: Activity-based discrimination and matching gains on Atanas. **(A)** Relationship between activity-based discrimination and Top-1 improvement. Each point represents one neuron identity. Discrimination is measured as the difference between similarity to the correct identity and the mean similarity to spatially nearby incorrect candidates. Larger differences are associated with greater improvements from adding activity. **(B)** Six examples in which activity corrects errors made using geometry alone. Spatially nearby candidate identities are difficult to distinguish from geometry, whereas their activity relationships to surrounding neurons provide additional information that allows NeuReL to recover the correct identities.

Table 3: Component ablations. Values are the unweighted mean ± sample SD of Top-1 accuracy across test–reference pairs (%).

Variant	Atanas	Kato
NeuReL	**78.06 ± 16.92**	**74.62 ± 26.85**
w/o Activity	46.75 ± 18.68	56.90 ± 33.02
w/o Geometry	42.13 ± 9.15	31.78 ± 25.03
w/o Population Encoder	64.93 ± 20.67	51.64 ± 27.87
w/o Relation Transport	75.82 ± 17.51	68.34 ± 28.91

Table 4: Controlled comparison of matching algorithms with fixed NeuReL embeddings and atlas identities.

Matcher	Atanas	Kato
Sinkhorn	64.18 ± 16.30	72.86 ± 26.87
FGW	53.26 ± 15.14	59.90 ± 32.45
FUGW	70.71 ± 15.67	73.05 ± 27.13
NeuReL	**78.06 ± 16.92**	**74.62 ± 26.85**

4.4 Ablations

We first evaluate the contribution of geometry and activity by retraining NeuReL without each modality (Table 3). Removing activity reduces Top-1 accuracy from 78.06% to 46.75% on Atanas and from 74.62% to 56.90% on Kato. Removing geometry reduces accuracy to 42.13% on Atanas and 31.78% on Kato, corresponding to larger drops of 35.93 and 42.84 points. These results support the complementary roles of geometry and activity, with their relative contributions varying across datasets. The contribution of activity is also visible in Figure 3: NeuReL correctly identifies all eight highlighted neurons in the enlarged region, compared with five for the model trained without activity.

We next evaluate the population encoder and relation transport by removing each component and retraining under the same protocol. Removing the population encoder reduces Top-1 accuracy to 64.93% on Atanas and 51.64% on Kato, corresponding to drops of 13.13 and 22.98 percentage points. Removing relation transport reduces accuracy to 75.82% and 68.34%, corresponding to drops of 2.24 and 6.28 points, respectively. These results support using population relations both to contextualize neuron representations and to refine correspondences.

We further compare matching algorithms without retraining, using fixed NeuReL node embeddings and atlas identities (Table 4). Sinkhorn achieves 64.18% and 72.86% Top-1 accuracy on Atanas and Kato, respectively. Incorporating structural consistency with FGW gives 53.26% and 59.90%, while FUGW reaches 70.71% and 73.05%. NeuReL achieves 78.06% and 74.62%, exceeding the strongest alternative, FUGW, by 7.35 percentage points on Atanas and 1.57 points on Kato. FGW and FUGW construct structural costs from distances between node embeddings, whereas NeuReL directly compares learned pairwise relation vectors.

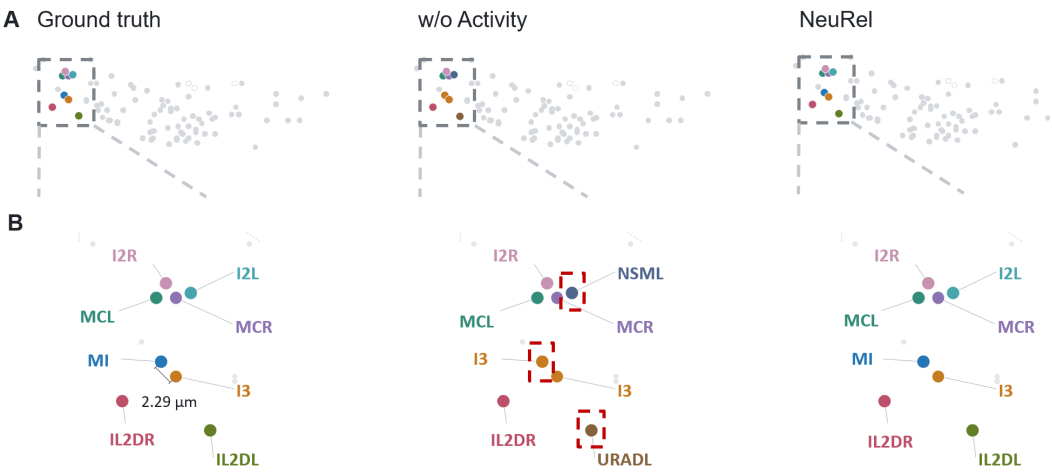

Figure 3: Contribution of neural activity to neuron matching in an Atanas recording. **(A)** Ground-truth identities, predictions without activity, and predictions from NeuReL for the same recording. **(B)** Enlarged views of the boxed region in (A), where red dashed boxes indicate incorrectly identified neurons. NeuReL correctly identifies all eight highlighted neurons, compared with five when activity is removed.

5 Conclusion

We proposed NeuReL, a supervised framework that matches neurons through their relations with the surrounding population, combining spatial positions and neural activity rather than relying on individual-neuron features alone. NeuReL learns geometric and functional relations and uses differentiable optimal transport to refine matches based on relational consistency across recordings. On held-out animals, NeuReL achieves the highest mean Top-1 accuracy among the evaluated methods on both *C. elegans* datasets and performs strongly in longitudinal zebrafish matching. Together with our ablation studies, these results demonstrate that combining multimodal data with population relations significantly improves neuron matching accuracy.

However, NeuReL has several limitations. First, it requires annotated neuron correspondences for supervised training, which may limit its use in datasets with scarce labels. Second, neural activity may not provide a stable identity signal, as activity patterns can overlap across neurons and vary with behavioral state, development, or recording conditions. Finally, our experiments demonstrate applicability across species, but do not establish zero-shot generalization to unseen species. Future work includes exploring unsupervised pretraining to reduce annotation requirements, developing domain adaptation methods to improve transfer to unseen species, and using neuron matching to integrate functional recordings with connectomic and molecular atlases.

AI USE STATEMENT

In this work, we used generative AI tools for improving the wording and readability of the manuscript and checking grammar. Generative AI tools were not used to develop research ideas, design methods or experiments, generate or analyse results, or make scientific claims. All AI-assisted edits were reviewed by the authors to ensure that they preserved the intended meaning and accurately reflected the work conducted. We take responsibility for the final content of this work, including all text and claims produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive personal data. All datasets used in this study are publicly available and were used in accordance with their respective terms and licenses. We are not aware of any significant ethical risks arising from the proposed methodology or its intended use. The authors declare no conflicts of interest relevant to this work.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the experiments throughout the main text and appendices. Appendix A reports additional controlled evaluations, rejection analysis, and sensitivity experiments. Appendix C describes the input preprocessing, model architecture, population relations, matching procedure, training objective, and hyperparameters. Appendix D specifies the dataset preparation, cross-validation splits, reference construction, evaluation metrics, and statistical aggregation. Appendix E documents the implementations and parameter selection procedures for all baseline methods. The datasets used in this work are publicly available and are cited in the main text. We use fixed data splits and report random seeds where applicable. Neural-network training uses the training split, while checkpoint selection, threshold calibration, and hyperparameter tuning use validation data. For zebrafish, StatAtlas and CRF_ID additionally use annotated auxiliary recordings from the held-out fish to construct reference models, excluding both evaluated endpoints. Endpoint identity annotations are used only for evaluation.

REFERENCES

Vinam Arora, Divyansha Lachi, Ian Knight, Mehdi Azabou, Blake A. Richards, Cole Hurwitz, Joshua H. Siegle, and Eva L. Dyer. Know thyself by knowing others: Learning neuron identity from population context. In *Advances in Neural Information Processing Systems*, volume 38, pp. 110311–110340, 2025. doi: 10.52202/085713-3325.

Adam A. Atanas, Jungsoo Kim, Ziyu Wang, Eric Bueno, McCoy Becker, Di Kang, Jungyeon Park, Talya S. Kramer, Flossie K. Wan, Saba Baskoylu, Ugur Dag, Elpiniki Kalogeropoulou, Matthew A. Gomes, Cassi Estrem, Netta Cohen, Vikash K. Mansinghka, and Steven W. Flavell. Brain-wide representations of behavior spanning multiple timescales and states in C. elegans. *Cell*, 186(19):4134–4151.e31, 2023. doi: 10.1016/j.cell.2023.07.035.

Shivesh Chaudhary, Sol Ah Lee, Yueyi Li, Dhaval S. Patel, and Hang Lu. Graphical-model framework for automated annotation of cell identities in dense cellular images. *eLife*, 10:e60321, 2021. doi: 10.7554/eLife.60321.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Raymond L. Dunn, Caitriona M. Costello, Jackson M. Borchardt, Daniel Y. Sprague, Grace C. Chiu, Julia M. Miller, Noelle D. L'Etoile, and Saul Kato. Short-term memory by distributed neural network oscillators in a simple nervous system. *Current Biology*, 35(22):5582–5593.e4, 2025. doi: 10.1016/j.cub.2025.10.018.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021. URL https://jmlr.org/papers/v22/20-451.html.

Ni Ji, Gurrein K. Madan, Guadalupe I. Fabre, Alyssa Dayan, Casey M. Baker, Talya S. Kramer, Ijeoma Nwabudike, and Steven W. Flavell. A neural circuit for flexible control of persistent behavioral states. *eLife*, 10:e62889, 2021. doi: 10.7554/eLife.62889.

Shunsuke Kamiya, Taiga Mitamura, Muneki Ikeda, and Masafumi Oizumi. Unsupervised neuronal matching with spontaneous neuronal activity. In *ICLR 2025 Workshop on Representational Alignment (Re-Align)*, 2025. URL https://openreview.net/forum?id=qAgQqVVwq9. Workshop poster.

Philipp J. Keller. Longitudinal functional imaging data for the zebrafish embryonic spinal cord. Janelia Research Campus, 2019. Dataset.

Enrico Lanza, Valeria Lucente, Martina Nicoletti, Silvia Schwartz, Ilaria F. Cavallo, Davide Caprini, Christopher W. Connor, Mashel Fatema A. Saifuddin, Julia M. Miller, Noelle D. L'Etoile, and Viola Folli. See Elegans: Simple-to-use, accurate, and automatic 3D detection of neural activity from densely packed neurons. *PLOS ONE*, 19(3):e0300628, 2024. doi: 10.1371/journal.pone.0300628.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017. doi: 10.1109/ICCV.2017.324.

Lu Mi, Trung Le, Tianxing He, Eli Shlizerman, and Uygar Sümbül. Learning time-invariant representations for individual neurons from population dynamics. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46007–46026, 2023. doi: 10.52202/075280-1994.

Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2262–2275, 2010. doi: 10.1109/TPAMI.2010.46.

Jeffrey P. Nguyen, Frederick B. Shipley, Ashley N. Linder, George S. Plummer, Mochi Liu, Sagar U. Setru, Joshua W. Shaevitz, and Andrew M. Leifer. Whole-brain calcium imaging with cellular resolution in freely behaving Caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 113(8):E1074–E1081, 2016. doi: 10.1073/pnas.1507110112.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2664–2672. PMLR, 2016. URL https://proceedings.mlr.press/v48/peyre16.html.

Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11143–11152, 2022. doi: 10.1109/CVPR52688.2022.01086.

Tina Schrödel, Robert Prevedel, Karin Aumayr, Manuel Zimmer, and Alipasha Vaziri. Brain-wide 3D imaging of neuronal activity in Caenorhabditis elegans with sculpted light. *Nature Methods*, 10(10):1013–1020, 2013. doi: 10.1038/nmeth.2637.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2074.

Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced gromov wasserstein. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21792–21804, 2022. doi: 10.52202/068431-1584.

Erdem Varol, Amin Nejatbakhsh, Ruoxi Sun, Gonzalo E. Mena, Eviatar Yemini, Oliver Hobert, and Liam Paninski. Statistical atlas of C. elegans neurons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 119–129. Springer, 2020. doi: 10.1007/978-3-030-59722-1_12.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020. doi: 10.3390/a13090212.

Vivek Venkatachalam, Ni Ji, Xian Wang, Christopher Clark, James Kameron Mitchell, Mason Klein, Christopher J. Tabone, Jeremy Florman, Hongfei Ji, Joel Greenwood, Andrew D. Chisholm, Jagan Srinivasan, Mark Alkema, Mei Zhen, and Aravinthan D. T. Samuel. Pan-neuronal imaging in roaming Caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 113(8):E1082–E1088, 2016. doi: 10.1073/pnas.1507109113.

Yinan Wan, Ziqiang Wei, Loren L. Looger, Minoru Koyama, Shaul Druckmann, and Philipp J. Keller. Single-cell reconstruction of emerging population activity in an entire developing circuit. *Cell*, 179(2):355–372.e23, 2019. doi: 10.1016/j.cell.2019.08.039.

J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 314(1165):1–340, 1986. doi: 10.1098/rstb.1986.0056.

Wei Wu, Can Liao, Zizhen Deng, Zhengrui Guo, and Jinzhuo Wang. Neuron platonic intrinsic representation from dynamics using contrastive learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/63403963ebabae8b5cc05e14b5761c4c-Abstract-Conference.html.

Yuxiang Wu, Shang Wu, Xin Wang, Chengtian Lang, Quanshi Zhang, Quan Wen, and Tianqi Xu. Rapid detection and recognition of whole brain activity in a freely behaving Caenorhabditis elegans. *PLOS Computational Biology*, 18(10):e1010594, 2022. doi: 10.1371/journal.pcbi.1010594.

Eviatar Yemini, Albert Lin, Amin Nejatbakhsh, Erdem Varol, Ruoxi Sun, Gonzalo E. Mena, Aravinthan D. T. Samuel, Liam Paninski, Vivek Venkatachalam, and Oliver Hobert. NeuroPAL: A multicolor atlas for whole-brain neuronal identification in C. elegans. *Cell*, 184(1):272–288.e11, 2021. doi: 10.1016/j.cell.2020.12.012.

Xinwei Yu, Matthew S. Creamer, Francesco Randi, Anuj K. Sharma, Scott W. Linderman, and Andrew M. Leifer. Fast deep neural correspondence for tracking and identifying neurons in C. elegans using semi-synthetic training. *eLife*, 10:e66410, 2021. doi: 10.7554/eLife.66410.

APPENDIX

A ADDITIONAL RESULTS

A.1 ACTIVITY RELATIONS DISTINGUISH SPATIALLY SIMILAR NEURONS

Figure 4 further examines whether relationships to surrounding neurons distinguish the correct identity from spatially nearby alternatives. For each held-out query identity, we compare the similarity of its geometric or activity relationships to the correct training identity and to nearby incorrect candidates. Geometric relationships provide little separation between the correct identity and nearby candidates, whereas activity relationships more clearly favor the correct identity. This provides additional evidence that activity complements geometry when spatial information alone is ambiguous.

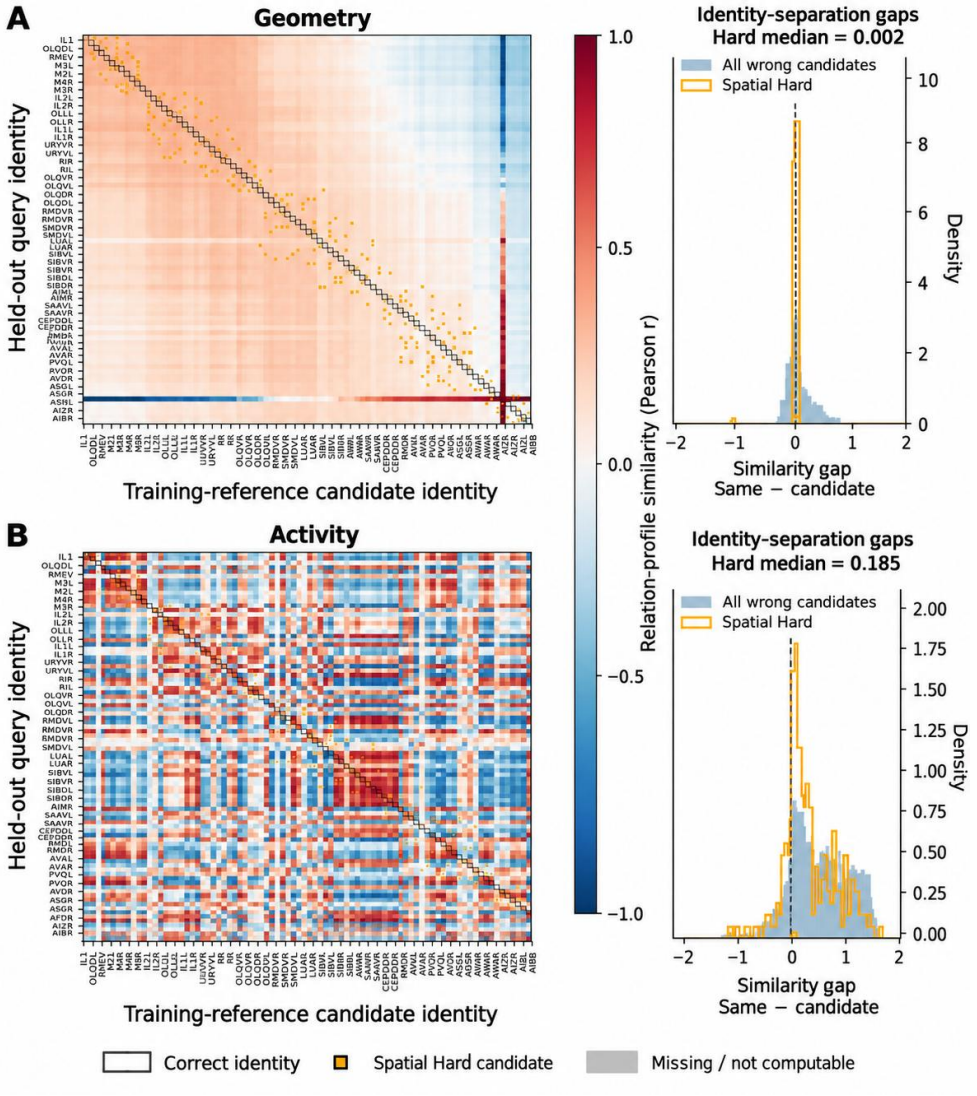

Figure 4: Relationships to surrounding neurons on Atanas. **A:** Similarity of geometric relationships between held-out query identities and training-reference candidates. **B:** Similarity of activity relationships for the same pairs. Black boxes indicate correct identities, orange markers denote spatially nearby incorrect candidates, and gray entries are missing or not computable. Right panels show the difference in similarity between the correct identity and alternative candidates. Activity relationships provide greater separation from nearby incorrect candidates than geometry alone.

A.2 Robustness to Query Perturbations

We evaluate NeuReL on Kato after perturbing query coordinates, removing query neurons, or adding synthetic distractors, while keeping the trained model and reference fixed. Figure 5 reports Top-1 on the same evaluation queries as the main benchmark, restricted to surviving original queries after removal. Mean accuracy is 74.62% on clean recordings, 44.79% at coordinate-noise SD 0.20, 65.78% after removing 50% of query neurons, and 68.03% after adding 50% distractors. These results characterize NeuReL's sensitivity to the tested perturbations; they do not compare robustness across methods.

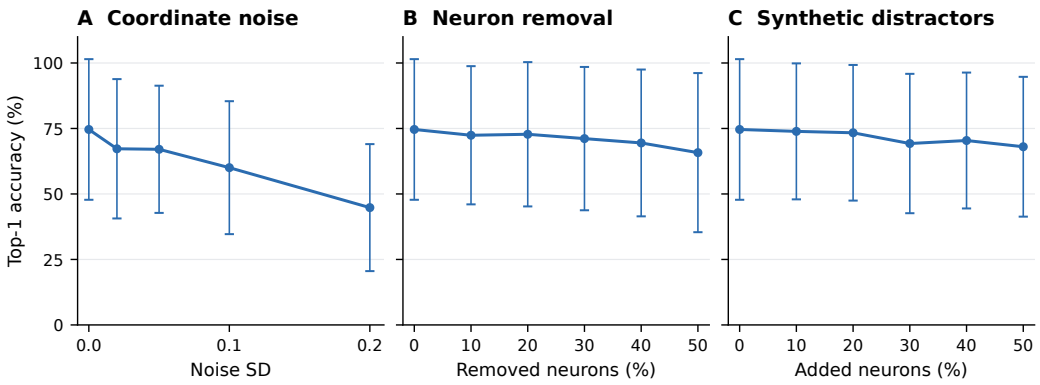

Figure 5: NeuReL sensitivity to query perturbations on Kato. (A) Coordinate noise in normalized coordinates. (B) Random removal of observed query neurons. (C) Synthetic distractors added relative to the original query population size. Points and error bars show the unweighted mean and sample SD across 87 records. Nonempty perturbation replicates are averaged within each record before this summary; the clean condition is evaluated once. Model parameters and references are fixed. Added distractors are excluded from the matching-accuracy denominator. Error bars show between-record variation, not confidence intervals.

Removing neurons can leave a recording–replicate combination without any eligible query. Such combinations have undefined accuracy and are excluded from the within-record average. All 87 records retain at least one nonempty replicate at every removal level. Table 5 gives the contributing counts; the clean condition contains 499 evaluation queries.

Table 5: Scoring coverage under query-neuron removal on Kato. There are 261 planned recording–replicate combinations per nonzero level. Query counts sum surviving evaluation queries across the three perturbation replicates.

Removed (%)	Nonempty recording–replicates	Surviving queries
10	260	1,338
20	258	1,195
30	253	1,065
40	248	864
50	241	736

A.3 Sensitivity to Test-Time Activity Duration

We vary the activity available at test time while holding the trained model, geometry, and atlas fixed. For each duration, we average accuracy over beginning, middle, and end windows within each record, then summarize records with equal weight. Short windows yield substantially lower accuracy on both datasets (Figure 6; Table 6). At 900 seconds, mean Top-1 reaches 77.05% on Atanas and 72.51% on Kato, respectively 1.01 and 2.11 percentage points below the complete-recording condition. This experiment measures test-time truncation of a fixed model, not the effect of training with different recording durations.

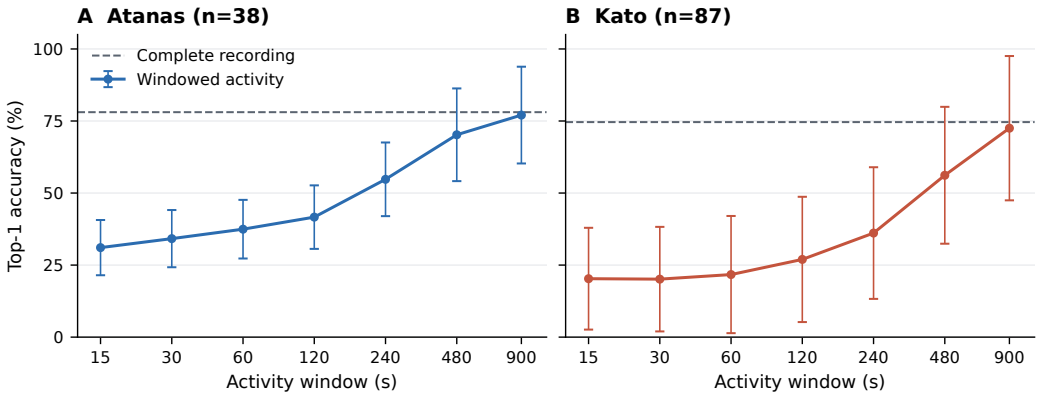

Figure 6: Sensitivity to test-time activity duration on the main benchmark evaluation queries. Each record contributes the mean Top-1 over its three window positions; points and error bars show the mean and sample SD across records (38 Atanas, 87 Kato). Dashed lines indicate complete-recording means under the same scoring rule. Window duration uses a logarithmic axis. The trained model, geometry, reference candidates, and atlas remain fixed.

Table 6: Test-time activity duration: Top-1 mean $\pm$ sample SD across records (%). The three positions are averaged within each record. Complete uses the full recorded activity.

Window (s)	Atanas	Kato
15	31.06 ± 9.58	20.28 ± 17.66
30	34.18 ± 9.93	20.12 ± 18.14
60	37.46 ± 10.17	21.71 ± 20.33
120	41.64 ± 11.02	26.97 ± 21.74
240	54.77 ± 12.78	36.11 ± 22.84
480	70.22 ± 16.08	56.17 ± 23.75
900	77.05 ± 16.79	72.51 ± 25.04
Complete	78.06 ± 16.92	74.62 ± 26.85

A.4 Unmatched Neuron Rejection

We evaluate rejection using the main model's frozen static atlas, with thresholds calibrated anew on validation recordings and fixed across all test distractor levels. Known neurons have unique, reliable annotations represented in this atlas; unknown neurons are the materialized synthetic distractors. Native unannotated, uncertain, duplicate, or reference-absent identities remain in the model input but are excluded from these two scoring groups. Known-neuron metrics use 93 Kato records, while nonzero distractor conditions contain unknown neurons in all 95 records.

Figure 7 shows the trade-off between recovering unknown neurons and rejecting known ones. At the validation recall target of 90%, increasing the distractor fraction from 10% to 50% changes test unknown recall from 84.08% to 89.19% and known false rejection from 22.32% to 30.81%. On clean recordings, the same fixed thresholds reject 20.37% of known neurons. At 50% distractors, Top-1 on known neurons is 60.86% before rejection and 53.46% after rejection. These values use all known neurons as their denominator; they are not accuracies restricted to accepted predictions.

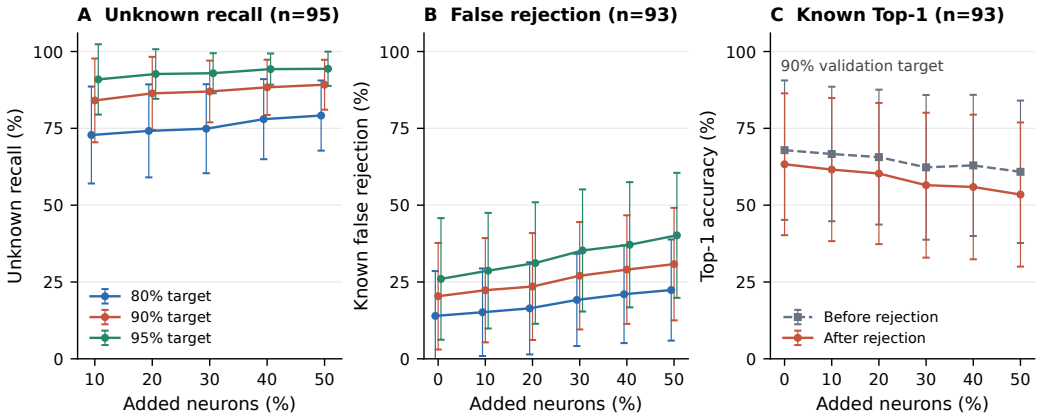

Figure 7: Rejection with the main static atlas on Kato. (A) Unknown recall and (B) known false rejection for log-margin thresholds calibrated to validation recall targets of 80%, 90%, and 95%. One threshold per fold and target is fixed across all test conditions. (C) Known-neuron Top-1 before and after rejection at the 90% validation target. Means and sample SD are computed across records after averaging perturbation replicates within each record. Known metrics use 93 records; unknown recall uses 95 records and is undefined for the clean condition.

Table 7: Kato rejection at the 90% validation recall target. Entries are mean ± sample SD across records (%). Known metrics use $N = 93$; unknown recall uses $N = 95$ at nonzero distractor fractions. The clean condition has no unknown neurons.

Added (%)	Unknown recall	Known false rejection	Top-1 before rejection	Top-1 after rejection
0	—	20.37 ± 17.34	67.91 ± 22.69	63.31 ± 23.08
10	84.08 ± 13.63	22.32 ± 16.97	66.65 ± 21.89	61.58 ± 23.28
20	86.38 ± 11.89	23.52 ± 17.40	65.64 ± 21.95	60.29 ± 22.98
30	87.00 ± 10.05	27.03 ± 17.48	62.30 ± 23.55	56.50 ± 23.58
40	88.35 ± 8.99	29.03 ± 17.67	62.93 ± 22.98	55.92 ± 23.52
50	89.19 ± 8.12	30.81 ± 18.31	60.86 ± 23.18	53.46 ± 23.45

A.5 Reference Choice

We compare a single training medoid with a population atlas using the same frozen seed42 encoder and the evaluation queries used in the main benchmark. Each reference retains its original candidate identities; an evaluated identity absent from the reference receives zero credit.

Reference choice affects matching accuracy even with the encoder and evaluation queries held fixed (Figure 8; Table 8). On Atanas, the population atlas and single medoid have similar mean Top-1 accuracies (78.06% and 77.29%, respectively). On Kato, the single medoid has higher mean accuracy than the population atlas (86.43% versus 74.62%). The population-atlas inference strategy was selected using Atanas validation data and then applied unchanged to Kato. The higher single-medoid result on Kato is therefore reported as a post hoc reference-choice sensitivity analysis and was not used to select the main Kato result.

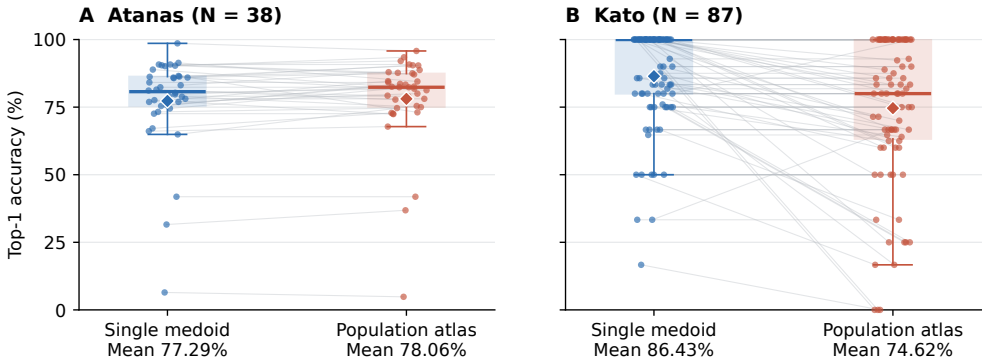

Figure 8: Reference choice for NeuReL on the main benchmark evaluation queries. Panels show Atanas (38 records) and Kato (87 records). Dots are individual records, connected across reference choices. Boxes show medians and interquartile ranges, with whiskers extending to the most extreme points within 1.5 interquartile ranges; diamonds denote means. Both reference variants use the same frozen encoder.

Table 8: Reference variants with a fixed NeuReL encoder and the main benchmark evaluation queries. Values are record-level means ± sample SD (%; 38 Atanas and 87 Kato records). Hungarian assignment is recomputed after selecting eligible query rows, with all original reference columns retained.

Dataset	Reference	Top-1	Hungarian
Atanas	Single medoid	77.29 ± 17.31	78.08 ± 16.88
Atanas	Population atlas	78.06 ± 16.92	78.27 ± 17.08
Kato	Single medoid	86.43 ± 18.28	88.18 ± 19.55
Kato	Population atlas	74.62 ± 26.85	74.67 ± 26.83

A.6 Controlled Comparison of Matching Algorithms

Table 4 compares matching algorithms with the NeuReL encoder, atlas node embeddings, and candidate identities held fixed. All controls use the same evaluation queries and pair-level aggregation as the main benchmark (Appendix D.3). NeuReL has the highest mean Top-1 accuracy on both datasets. FGW (Vayer et al., 2020) and FUGW (Thual et al., 2022) construct structural costs from distances between node embeddings, whereas NeuReL directly compares learned relation vectors. The comparison therefore tests both the structural inputs and the matching updates, rather than isolating the update algorithm alone. Appendix E.5 gives the costs and validation grids.

A.7 Coverage of Atlas Relations

Identity pairs never jointly annotated account for 8.43% of directed atlas entries on Atanas and 55.97% on Kato, counting diagonals and pooling entries across folds. NeuReL assigns zero vectors to these entries and includes them in the relation discrepancy, as specified in the main method. These percentages describe annotation coverage of the training atlases.

B Additional Relation Analyses

B.1 Relationships to Surrounding Neurons Across Animals

For each ordered pair of Atanas animals, we compare each neuron through its relationships to shared annotated anchor neurons, excluding the identities of the two neurons being compared. Geometric relationships are represented by negative Euclidean distances after independent centering and scaling, whereas activity relationships are represented by within-recording Pearson correlations. For the multimodal analysis, separately standardized geometric and activity relationships are concatenated. Similarity between two neurons is computed as the Pearson correlation between their aligned relationship vectors.

For each query identity, we compare the correct identity with three incorrect identities nearest to its position in the target animal. We quantify discrimination as the similarity to the correct identity minus the mean similarity to these nearby incorrect candidates, averaged across animal pairs for each identity. Geometry has mean correct-identity similarity 0.864 and a mean difference of 0.023, compared with 0.363 and 0.257 for activity, and 0.614 and 0.140 for their combination. Thus, activity relationships are less consistent across animals overall, but better distinguish the correct neuron from spatially nearby alternatives.

B.2 Relation Distinctiveness and Matching Accuracy

We next ask whether neurons whose relationships more clearly distinguish the correct identity from nearby incorrect candidates are also matched more accurately by NeuReL. Using frozen seed42 population-atlas predictions on Atanas, we pool Top-1 accuracy by identity across 3,210 held-out queries. Each of the 149 identities with both a relationship-based discrimination measure and an accuracy estimate contributes one point. One identity lacking a current accuracy estimate is excluded only from this correlation analysis, without changing the benchmark query sets. This is separate from the activity-gain analysis in Section 4.3, whose response is the improvement from adding activity rather than absolute accuracy.

Geometric margins have Spearman correlation $\rho = 0.165$ with absolute accuracy, compared with $\rho = 0.339$ for activity and $\rho = 0.346$ for multimodal margins (Figure 9). These are descriptive associations across related neuronal identities; they neither isolate a causal contribution of a modality nor establish out-of-sample predictive performance. Identity-level aggregation here is distinct from the record-level averages in the main benchmark.

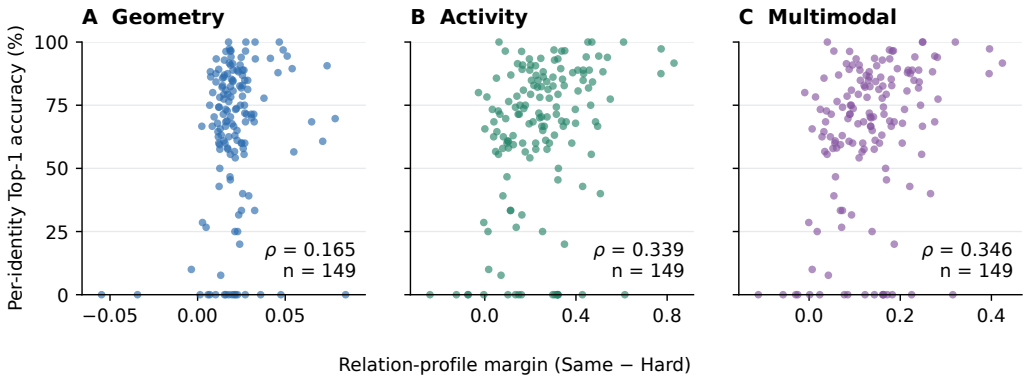

Figure 9: Relationship distinctiveness and absolute NeuReL matching accuracy on Atanas. Each point represents one of 149 neuron identities. Horizontal axes show the difference between similarity to the correct identity and mean similarity to nearby incorrect candidates; the vertical axis shows identity-level Top-1 accuracy from the frozen seed42 model. The three panels report descriptive Spearman correlations for geometry, activity, and their combination.

B.3 Cross-Animal Consistency of Population Relations

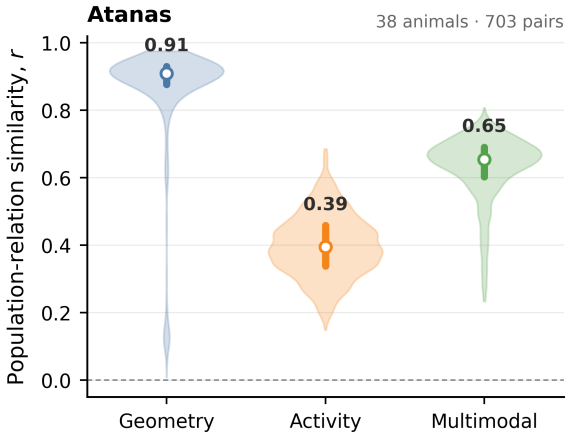

Figure 10: Population-level relation similarity on Atanas across 703 unordered pairs of 38 animals, using shared annotated identities. Open circles show medians and vertical bars show interquartile ranges.

We align the shared identities of each unordered animal pair and correlate the off-diagonal upper-triangular entries of their relation matrices. Here, the multimodal matrix averages separately standardized geometry and activity entries rather than concatenating profiles. Across 703 pairs of 38 animals, geometric relationships are the most consistent across animals, whereas activity relationships are less consistent overall (Fig. 10). Together with the identity-level analysis above, this shows that cross-animal consistency and the ability to distinguish nearby neurons are different properties: activity can vary more across animals while still providing stronger discrimination between spatially similar candidates.

C Additional Method Details

C.1 Population Sampling and Input Preparation

Training pairs share at least two unique, valid identities for *C. elegans* and twenty for zebrafish. We shuffle pairs each epoch and process one pair per optimization step without population padding. Neurons with missing or duplicate annotations contribute context but no supervised targets.

Known correspondences are established before independently removing neurons from each recording with probability 0.05, retaining at least two when available. Surviving pairs receive bidirectional match targets; a surviving neuron whose established counterpart was removed receives an unmatched target. Missing annotations alone never supply unmatched supervision.

For the *C. elegans* datasets, although neuronal positions may vary over time in the original recordings, NeuReL uses a single three-dimensional coordinate for each neuron from the processed data rather than its positional trajectory. We exclude invalid coordinates, median-center each axis, and divide by its RMS deviation from the median, floored at 10^{-5}. Nonfinite activity samples are replaced by their trace median, or zero if the trace has no finite samples. Traces are linearly resampled to 512 points for *C. elegans* and 128 for zebrafish, then standardized by temporal mean and SD, with SD floored at 10^{-5}. Summary statistics are computed before standardization.

C.2 Model Architecture

Encoders and fusion. Table 9 specifies the encoders. For embeddings $\mathbf{g}, \mathbf{a} \in \mathbb{R}^D$, let $\mathbf{v} = [\mathbf{g}; \mathbf{a}; \mathbf{g} \odot \mathbf{a}]$. A $3D \to D \to 2$ MLP followed by softmax gives modality weights (w_g, w_a), and a $3D \to 2D \to D$ MLP supplies a residual

$$\mathrm{Fuse}(\mathbf{g}, \mathbf{a}) = \mathrm{LN}(w_g \mathbf{g} + w_a \mathbf{a} + F_{\mathrm{res}}(\mathbf{v})) \,. \tag{6}$$

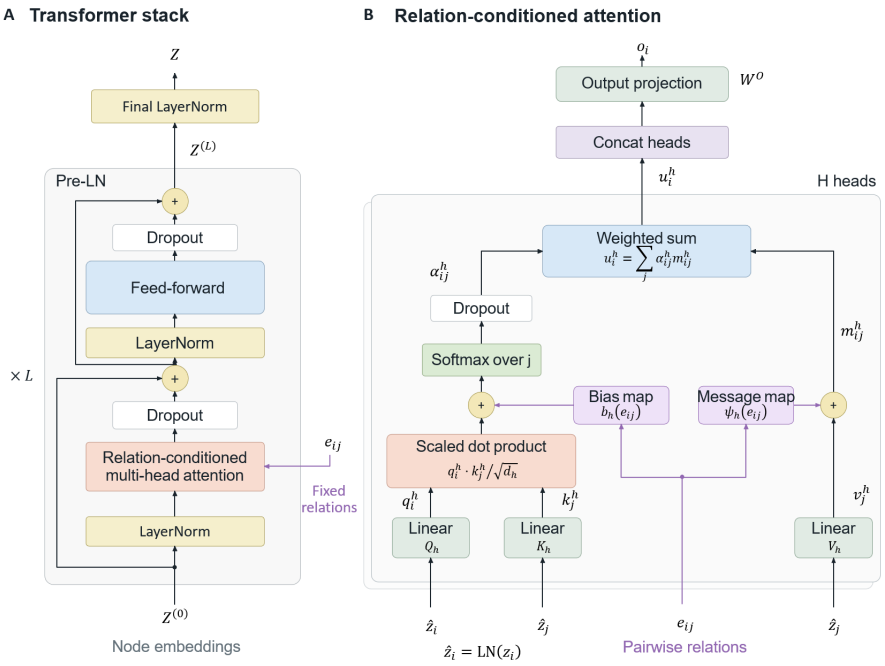

Figure 11: Relation-conditioned Transformer. (A) Pre-LayerNorm attention and feed-forward sublayers use residual connections; a final LayerNorm follows the stack. (B) Pairwise embeddings provide attention biases and messages added to projected values. Head outputs are concatenated and projected. Pairwise embeddings are fixed across layers, whose parameters are independent.

Both MLPs use GELU. The residual uses dropout and a zero-initialized final linear layer. Node and pairwise fusion have separate parameters, with $D = 96$ and 48, respectively.

Table 9: NeuReL encoder architecture. MLP widths include input and output dimensions. Dropout is 0.10 wherever used; LN denotes LayerNorm.

Component	Configuration
Geometry encoder	Input $[\mathbf{x}_i; \|\mathbf{x}_i\|_2]$; MLP $4 \to 96 \to 96$, hidden GELU and dropout, final LN.
Activity CNN	Standardized trace and first difference (zero prepended; not separately standardized); three Conv1D layers with 32 channels, kernels $9, 7, 5$, unit stride, length-preserving padding, and GELU. Temporal mean/max pooling gives 64 features.
Activity summaries	Pre-standardization mean, SD, and 25th/50th/75th percentiles; MLP $5 \to 48 \to 48$ with hidden GELU.
Activity projection	Concatenated CNN and summary features; MLP $112 \to 96 \to 96$, hidden GELU and dropout, final LN.
Pairwise encoders	Geometry $8 \to 48 \to 48$; activity $4 \to 48 \to 48$. Both use hidden GELU and final LN, followed by pairwise fusion.
Population encoder	Two pre-LN Transformer layers, four heads, node width 96; feed-forward width $96 \to 384 \to 96$. Final LN yields $\mathbf{z}_i$.
Relation head	Input $[\mathbf{e}_{ij}; \mathbf{z}_i - \mathbf{z}_j; \mathbf{z}_i \odot \mathbf{z}_j]$; MLP $240 \to 96 \to 8$, hidden GELU and final $\tanh$.

Pairwise descriptors. For every ordered pair, including self-pairs, we use

$$\mathbf{g}_{ij}^{x} = \left[\mathbf{x}_i - \mathbf{x}_j; d_{ij}; \left\{\exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right)\right\}_{\sigma\in\mathcal{S}}\right], \qquad \mathbf{g}_{ij}^{a} = \left[c_{ij}; c_{ij}^{\Delta}; [c_{ij}]_+; [-c_{ij}]_+\right], \tag{7}$$

where $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, $\mathcal{S} = \{0.25, 0.5, 1, 2\}$, and $[v]_+ = \max(v, 0)$. The terms c_{ij} and c_{ij}^{Δ} are within-recording Pearson correlations of traces and separately standardized, unpadded first differences. Geometric displacement makes the fused embedding $\mathbf{e}_{ij}$ directional; activity correlations are symmetric.

Population encoding. Each Transformer layer implements Eq. 1 with separate bias-free projections of $\mathbf{e}_{ij}$ into attention biases and value messages (Fig. 11). Head outputs are concatenated and projected. Dropout is applied to attention weights, projected attention outputs, and feed-forward hidden and output features. Pairwise embeddings remain fixed across layers; layer parameters are independent.

C.3 Matching Implementation

Partial transport. For n query and m reference neurons, separate learned logits b_Q, b_R supply the unmatched column and row, with a zero corner entry. We use augmented marginals

$$\widetilde{\boldsymbol{\mu}} = \frac{[\mathbf{1}_n; m]}{n+m}, \qquad \widetilde{\boldsymbol{\nu}} = \frac{[\mathbf{1}_m; n]}{n+m}. \tag{8}$$

Temperatures are $\tau_u = 0.01 + \mathrm{softplus}(\theta_u)$ and $\tau_r = 0.01 + \mathrm{softplus}(\theta_r)$, with $\lambda = \mathrm{softplus}(\theta_\lambda)$ and $\eta = \lambda/\tau_r$. Initial values and iteration counts appear in Table 10. Each log-domain Sinkhorn iteration performs one row and one column update.

Relation contraction. The real block $\mathbf{P}^{(t)}$ is normalized independently of unmatched mass

$$\bar{\mathbf{P}}^{(t)} = \frac{\mathbf{P}^{(t)}}{\max\{\sum_{k,l} P_{kl}^{(t)}, 10^{-8}\}}. \tag{9}$$

To avoid a four-index tensor, define $A_{ik} = \|\mathbf{r}_{ik}^Q\|_2^2$, $B_{jl} = \|\mathbf{r}_{jl}^R\|_2^2$, $\mathbf{a}^{(t)} = \bar{\mathbf{P}}^{(t)}\mathbf{1}_m$, and $\mathbf{b}^{(t)} = (\bar{\mathbf{P}}^{(t)})^\top \mathbf{1}_n$. With relation-channel matrices $\mathbf{R}_{Q,d} \in \mathbb{R}^{n\times n}$ and $\mathbf{R}_{R,d} \in \mathbb{R}^{m\times m}$, the discrepancy is

$$\mathbf{D}^{(t)} = (\mathbf{A}\mathbf{a}^{(t)})\mathbf{1}_m^\top + \mathbf{1}_n(\mathbf{B}\mathbf{b}^{(t)})^\top - 2\sum_{d=1}^{8} \mathbf{R}_{Q,d}\bar{\mathbf{P}}^{(t)}\mathbf{R}_{R,d}^\top. \tag{10}$$

This identity holds for directed relations. Negative roundoff is clamped to zero. Each refinement subtracts $\eta\mathbf{D}^{(t)}$ from the original node logits, rather than accumulating discrepancies. Gradients pass through the plan, contraction, and Sinkhorn updates. Each solve uses a fixed current score matrix; the learned refinements do not assume minimization of one global FGW objective.

Algorithm 1 Relation-aware neuron matching

Require: Query $(\mathbf{Z}_Q, \mathbf{R}_Q)$, reference $(\mathbf{Z}_R, \mathbf{R}_R)$; learned τ_u, η, b_Q, b_R
1: Normalize each node embedding to unit length, giving $\bar{\mathbf{Z}}_Q, \bar{\mathbf{Z}}_R$
2: $\mathbf{L}^{\mathrm{node}} \leftarrow \bar{\mathbf{Z}}_Q\bar{\mathbf{Z}}_R^\top/\tau_u$
3: Set marginals by Eq. 8
4: $\mathrm{Aug}(\mathbf{L}) \leftarrow \begin{bmatrix} \mathbf{L} & b_Q\mathbf{1}_n \\ b_R\mathbf{1}_m^\top & 0 \end{bmatrix}$
5: $\widetilde{\mathbf{P}}^{(0)} \leftarrow \mathrm{Sinkhorn}_{20}(\mathrm{Aug}(\mathbf{L}^{\mathrm{node}}); \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\nu}})$
6: **for** $t = 0, 1$ **do**
7: Extract and normalize the real block using Eq. 9
8: Compute $\mathbf{D}^{(t)}$ using Eq. 10 and clamp to nonnegative values
9: $\widetilde{\mathbf{P}}^{(t+1)} \leftarrow \mathrm{Sinkhorn}_{20}(\mathrm{Aug}(\mathbf{L}^{\mathrm{node}} - \eta\mathbf{D}^{(t)}); \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\nu}})$
10: **end for**
11: **return** $\widetilde{\mathbf{P}}^{(2)}$

Directional supervision. We divide each final matching weight by its prescribed source marginal

$$p_{Q\to R}(j \mid i) = \frac{\widetilde{P}_{\theta,ij}^{(T)}}{\widetilde{\mu}_i}, \qquad p_{R\to Q}(i \mid j) = \frac{\widetilde{P}_{\theta,ij}^{(T)}}{\widetilde{\nu}_j}. \tag{11}$$

Sources are real neurons and targets include the unmatched state. Finite Sinkhorn iterations only approximately satisfy the marginals. We clip supervised target scores to $[10^{-8}, 1]$, apply the focal loss in Eq. 4 with $\gamma = 2$, and average over supervised neurons in both directions.

C.4 Training and Dataset-Specific Inference

Table 10: NeuReL optimization and matching settings. The two refinements follow one initial solve, giving three Sinkhorn solves in total.

Parameter	Value
Optimizer / epochs / seed	AdamW / 80 / 42
Learning rate / weight decay	Constant 2×10^{-4} / 10^{-4}
Gradient-norm clipping / dropout	1.0 / 0.10
Batch / neuron-removal probability	One recording pair / 0.05
Node / pairwise / final relation dimensions	96 / 48 / 8
Activity length (worm / zebrafish)	512 / 128
Sinkhorn iterations / refinement steps	20 / 2
Initial τ_u / τ_r / λ	0.10 / 0.50 / 1.0
Initial unmatched logits / corner	-1.0 / 0 (fixed)
Focal exponent	2
Checkpoint selection	Validation bidirectional Top-1, every epoch

Checkpoint selection. Validation uses real-candidate Top-1 without an atlas. For worms, we pool known correspondences in both directions over all unordered validation recording pairs sharing at least two unique, valid identities. For zebrafish, we use both directions of the designated pairs from the validation fish (Appendix D.1). The highest-scoring checkpoint is frozen before reference construction.

Population atlas. For worms, the frozen encoder processes training recordings in evaluation mode, using all observed neurons as context. Each identity prototype averages its unit-normalized node embeddings and is normalized again. Only unique, valid annotations contribute. Directed relations are averaged over recordings jointly annotating both identities; unobserved pairs receive zero vectors and remain included in the discrepancy. The resulting atlas is fixed during test inference. The choice of direct atlas inference was made using Atanas validation data and then held fixed; this is separate from checkpoint selection, which uses recording pairs without an atlas.

Longitudinal matching. For zebrafish, the frozen model independently encodes and directly matches the two designated endpoints. NeuReL uses no population identity atlas, and endpoint identities are used only for evaluation. Ranking and Hungarian decoding follow Appendix D.3.

C.5 Computational Cost

Model size. Table 11 compares the neural-network parameter counts of NeuReL and the evaluated neural baselines. We count all network weights, including frozen parameters, but exclude atlas statistics, training-only projection heads, and per-instance optimization variables. NeuReL contains 0.437 M parameters, fewer than the compared neural baselines.

Inference time. We benchmark NeuReL on a workstation equipped with two NVIDIA GeForce RTX 4090 GPUs. Latency includes input preprocessing, population encoding, matching, and Hungarian assignment over the complete query population. Disk I/O, model loading, and one-time atlas encoding are excluded. We report the mean and sample standard deviation over 30 timed runs after warm-up.

Table 11: Neural-network parameter counts of the evaluated configurations. M denotes 10^6 parameters.

Method	Parameters (M)
fDNC	4.152
GeoTransformer	5.210
NuCLR	6.319
NEUREL (Ours)	**0.437**

Table 12: CPU inference time of NEUREL, including Hungarian assignment. Values are the mean $\pm$ sample SD over 30 repetitions. N and M denote the numbers of query neurons and atlas candidates, respectively.

Dataset	N	M	Time (ms)
Atanas	109	157	31.69 ± 4.37
Kato/RLD	71	145	20.15 ± 5.66

D DATASETS AND EVALUATION PROTOCOLS

D.1 DATASET PREPARATION AND SPLITS

Table 13 summarizes the datasets and imaging conditions used in our experiments.

***C. elegans*.** We use 38 Atanas and 95 Kato animals, with three-dimensional coordinates and calcium traces. Atanas recordings contain 109–153 observed neurons and 51–111 unique, valid identities; Kato recordings contain 45–140 neurons and 0–53 such identities. Atanas uses Dandiset 000776, version 0.241009.1509. Other annotations contribute context but no correspondence targets.

All recordings from an animal remain in one split. Five grouped partitions use split seed 20260823, keyed by Atanas recording identifier and Kato subject identifier. Fold f tests partition f, validates on $(f+1) \bmod 5$, and trains on the remaining three. Each animal is tested once (Table 14). Benchmark neural-network training uses seed 42 unless stated otherwise.

Zebrafish. Eight fish contain 61–122 tracked neurons each. In each leave-one-fish-out fold, one fish is tested, the next in a fixed cyclic order is validated, and six are used for training. Validation and test pairs use 30-second windows with start times spaced by 15 minutes and query/reference starts separated by 60 minutes. Only endpoints with the same `pair_id` are matched, in both directions. The eight test folds contain 101 physical pairs. Tracking identities define within-fish correspondences, with no shared canonical identities across fish.

D.2 REFERENCE SELECTION

Worm references. Single-reference methods use one training medoid per fold. After median-centering and dividing coordinates by 200, we compute the symmetric average of the two directed mean nearest-neighbor distances between recordings. The medoid minimizes mean distance to other training recordings; ties use recording identifier. Selection uses geometry only, after which each method applies its own preprocessing. NeuReL builds a population atlas from training recordings. StatAtlas uses a training atlas to construct posterior-overlap scores against the fixed training medoid. CRF_ID uses its official template with a coordinate-frame transform estimated from training animals. All references are fixed at test time.

Zebrafish references. Direct methods match the designated endpoints. StatAtlas and CRF_ID instead use 8–20 auxiliary time points from the same held-out fish, excluding both evaluated endpoints. Tracking identities link these auxiliary recordings, providing additional annotated reference information to these two baselines. Endpoint identities are reserved for scoring; all methods use the same test pairs.

Table 13: Datasets used in our experiments. Counts are reported per recording for *C. elegans* and per fish for zebrafish.

Dataset	Organism	Imaging condition	Subjects	Neurons	IDs
Atanas	*C. elegans*	Freely moving	38	109–153	51–111
Kato	*C. elegans*	Stimulation	95	45–140	0–53
Zebrafish	Zebrafish	Immobilized	8	61–122	61–122

Table 14: Animal counts by split. A range of fold indices indicates the same counts in each listed fold.

Dataset	Fold	Train	Validation	Test
Atanas	0	22	8	8
Atanas	1	22	8	8
Atanas	2	23	7	8
Atanas	3	24	7	7
Atanas	4	23	8	7
Kato	0–4	57	19	19
Zebrafish	1–8	6	1	1

D.3 Metrics and Aggregation

Evaluation queries. For the main matching benchmark, ablations, and fixed-encoder matcher controls on the two *C. elegans* datasets, we score query neurons with unique, valid annotations whose identities occur in the fixed training medoid for that fold. The same query subset is used for every method, including atlas methods, while each method retains its own candidate identities. An eligible identity absent from a method's candidates remains in the denominator and receives zero Top-k, reciprocal-rank, and Hungarian credit. Each held-out recording contributes one query–reference pair; recordings with no eligible queries are excluded, leaving 38 pairs on Atanas and 87 on Kato. Zebrafish scoring uses unique, valid tracking identities present at both endpoints. Unlabeled or ambiguous neurons provide population context but no evaluation targets.

Ranking and assignment. Top-1 uses independent argmax over real reference candidates, resolving ties by the original candidate-column order. Top-k and MRR use a stable descending ordering with the same tie rule. For Hungarian decoding, we first select the final eligible query rows and retain all original real-candidate columns, then solve the maximum-score one-to-one assignment. NeuReL uses the corresponding entries of the final real transport block without recomputing or renormalizing the plan after row selection. Complete neuronal populations remain available during encoding and matching; query filtering is applied only for evaluation. Missing predictions or absent correct candidates receive zero credit and remain in the denominator. The unmatched state is excluded from both ranking and Hungarian assignment.

Pair-level aggregation. For the main matching benchmark, ablations, and fixed-encoder matcher controls, we first compute each metric within a test–reference pair. For zebrafish, predictions in both directions are combined within each physical pair before computing its metric, so the 101 physical pairs are not counted as 202 independent observations. Given a metric value M_r for pair r, the reported mean and sample SD are

$$\overline{M} = \frac{1}{R}\sum_{r=1}^{R} M_r, \qquad s_M = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(M_r - \overline{M})^2}, \tag{12}$$

with $R = 38$, 87, and 101 for Atanas, Kato, and zebrafish, respectively. Each pair receives equal weight regardless of its number of eligible neurons. Accuracies are percentages and MRR lies in $[0, 1]$. Cross-validation determines training, validation, and test membership; the reported test summaries aggregate pairs rather than fold-level averages. Validation checkpoint selection follows the separate pooling procedure in Appendix C.4.

D.4 Supplementary Evaluation Protocols

Frozen models and references. The reference, activity-window, and query-perturbation analyses use the same frozen seed42 models as the main benchmark, with condition settings fixed before evaluation. Models are neither retrained nor selected by test performance. Each condition preserves its complete observed input population and original reference candidates; evaluation filters do not alter model inputs. Activity-window and perturbation results report real-candidate Top-1. Rejection uses the main model's frozen static atlas and thresholds calibrated only on validation data.

Activity windows. We extract windows of $15, 30, 60, 120, 240, 480$, and 900 seconds from the beginning, middle, and end of each test recording, resampling each to 512 points. Geometry and the atlas are unchanged. For each duration, the three positions receive equal weight within a record; the final mean and sample SD are computed across records. The complete-recording condition uses the same evaluation queries and reference as the main benchmark. Positions are not independent animals.

Query perturbations. We reuse the materialized Kato condition manifest for coordinate-noise SDs $0, 0.02, 0.05, 0.10, 0.20$ and removal or distractor fractions $0, 0.10, 0.20, 0.30, 0.40, 0.50$. Only queries are perturbed. Synthetic distractor coordinates are uniform in the original bounding box expanded on each axis by 0.25 times the median pairwise distance. Each distractor copies a uniformly sampled original trace, with replacement, and adds independent noise with SD 0.25 times the population activity SD at each time point, floored at 10^{-6} before scaling. The number added is $\max\{1, \mathrm{round}(rN)\}$ for nonzero fraction r and original population size N. Preprocessing and relations are recomputed for the perturbed input. The original candidate columns remain fixed.

Each nonzero condition uses three fixed perturbation replicates; the clean condition is evaluated once. We average available nonempty replicate accuracies within a record, then compute the unweighted record mean and sample SD. Removed original queries are excluded only when they are physically absent; other original eligible queries remain in the denominator. Added distractors supply context but no identity-ranking targets. Empty recording–replicate denominators are undefined, not zero; Table 5 reports their effect on scoring coverage.

Rejection targets and thresholds. For rejection, known neurons have unique, reliable annotations in the frozen atlas vocabulary. Unknowns are only the materialized synthetic distractors. Other native neurons provide context but are excluded from the binary rejection groups. Both before- and after-rejection accuracies use all known neurons as their denominator.

We use the log-margin score

$$s_i = \log p(\perp \mid i) - \max_{j \in \mathcal{C}_i} \log p(j \mid i), \tag{13}$$

where $\mathcal{C}_i$ contains all real candidates. Each fold pools validation scores from five nonzero distractor fractions and three perturbation replicates. For target recall $R_0 \in \{0.80, 0.90, 0.95\}$, let b be the $\lceil R_0 n_U \rceil$th largest score among the n_U synthetic unknowns. We set

$$\tau = \mathrm{nextafter}(b, -\infty), \qquad \delta_i = \mathbf{1}[s_i > \tau]. \tag{14}$$

This strict comparison includes ties at b, so validation recall can exceed its target. The threshold is fixed for all test recordings and severities in that fold. No test score is used for calibration.

Unknown recall and known false rejection use their respective group denominators; clean recordings have no defined unknown recall. Known-neuron Top-1 after rejection counts rejected known neurons as errors. At each distractor fraction we average perturbation replicates within a record, then take the mean and sample SD across records with a nonempty metric denominator. These recording-level summaries are kept separate from historical five-fold summaries; neither folds nor perturbation replicates replace the record as the reported averaging unit.

E Baseline and Matcher Implementations

All methods use the splits and evaluation queries in Appendix D. Training uses seed 42 and training data only; checkpoints and hyperparameters use validation data. Worm single-reference methods

share the fixed training medoid. Zebrafish methods directly match designated endpoints except for the auxiliary-reference access of StatAtlas and CRF_ID described above.

E.1 Geometry-Based Matching Baselines

CPD. For worms, coordinates are median-centered and divided by 200. We estimate in-plane rotation and three-dimensional translation at a fixed scale, apply Gaussian-kernel nonrigid CPD, and score correspondences using negative Euclidean distances. For zebrafish, deformable PyCPD median-centers and scales each coordinate axis by its RMS deviation, then registers both directions. Rankings use negative squared distances; Hungarian assignment uses the average of directional score matrices after transposing one.

fDNC. Worm fDNC starts from released pretrained weights with median-centered coordinates divided by 200. We fine-tune the last one or two Transformer layers and outlier head, using official matching/outlier losses and regularization toward pretrained weights. Zebrafish models are trained from scratch in each LOFO fold, median-centering and scaling each coordinate axis by its RMS deviation. Table 15 gives their separate settings. Dot-product scores follow joint query/reference contextualization; real-candidate ranking excludes the outlier output.

GeoTransformer. Coordinates are median-centered and divided by the 90th percentile of radial distances. The official KPConv pyramid, geometric Transformer, and coarse-to-fine matcher use coordinates and constant input features. Identities common to both recordings supervise patch overlap and fine matching/unmatched targets. Adam minimizes coarse circle and fine negative log-likelihood losses for 20,000 updates, with a learning rate 10^{-4}, warm-up, and cosine decay. Validation Top-1 selects the checkpoint. The maximum patch-level log transport score for each neuron pair supplies inference scores.

E.2 Atlas Baselines

StatAtlas. For worms, we retain the previously fixed alignment versions: legacy alignment on Atanas and PCA-permutation alignment on Kato. The frozen Gaussian atlas uses training annotations and regularized coordinate covariances. Query and fixed-medoid neurons are represented by equal-prior posteriors over atlas identities, and posterior overlap supplies pair scores over the original medoid candidate columns. All observed query neurons enter unlabeled alignment; test identities do not filter candidate columns or select an alignment version. Evaluated identities absent from the candidates receive zero credit.

For zebrafish, the auxiliary atlas excludes both evaluated endpoints. Independent unlabeled alignment gives identity posterior matrices $\mathbf{P}_q$ and $\mathbf{P}_r$ over shared atlas slots. We use

$$\mathbf{S} = \mathbf{P}_q \mathbf{P}_r^\top, \qquad S_{ij} = \sum_k p(k \mid q_i) p(k \mid r_j) \tag{15}$$

as pairwise correspondence scores. Endpoint identities supply evaluation targets only.

CRF_ID. We use angle-based relations, uniform unary potentials, UGM loopy belief propagation, and iterative duplicate-assignment resolution. Worm recordings are mean-centered and divided by RMS radius. A coordinate-frame transform is estimated from training identities shared with the official template and fixed during testing. Zebrafish uses the same auxiliary recordings as StatAtlas, excludes both endpoints, and reuses the geometry-only MATLAB inference adapter. Exported identity beliefs supply ranking and Hungarian scores.

E.3 NuCLR

We use the authors' calcium-imaging implementation as an activity-only baseline. Worm training pairs two temporal views of the same neuron within a recording, without canonical cross-animal identity supervision. Standardized 30-second views use decoupled contrastive loss. At inference, we average backbone embeddings over consecutive complete 30-second windows, normalize them, and match by cosine similarity with the projector disabled.

Zebrafish uses `NuclrV2gaCa2` with 128-frame windows, dimension 256, 32 latents, patch size 4, four heads, two temporal layers, and two spatiotemporal layers. Its training configuration is independent of the worm configuration (Table 15). Matching uses backbone cosine similarity with the projector disabled.

Table 15: Dataset-specific fDNC and NuCLR training. Zebrafish settings do not inherit worm configurations. All runs use seed 42.

Method/data	**Training and checkpoint selection**
fDNC / worms	Pretrained fine-tuning, AdamW, at most 60 epochs. Backbone learning rate $\{5 \times 10^{-6}, 10^{-5}\}$; select validation Top-1, with MRR breaking ties.
fDNC / zebrafish	Train from scratch in every LOFO fold; hidden dimension 128, six Transformer layers; at most 120 epochs, learning rate 2×10^{-4}, gradient clipping 1.0. Evaluate every two epochs on the validation fish, select ranking Top-1, and stop after 15 evaluations without improvement.
NuCLR / worms	Train from scratch, 30-second views, batch 16, temperature 0.2; 50,000 AdamW updates, maximum learning rate 1.25×10^{-4}, warm-up and cosine decay. Select validation Top-1, with MRR breaking ties.
NuCLR / zebrafish	30 epochs, 200 steps/epoch, four fish/batch, two temporal pairs/fish; learning rate 1.25×10^{-4}, weight decay 0.01, temperature 0.2, BF16. Select minimum validation SSL loss.

E.4 Optimal-Transport Baselines

FUGW. We use `fugw==0.1.1` with the dense solver, geometry for linear costs, activity distances for structure, and unbalanced marginals. Worm coordinates use training-set standardization. We first fill missing activity values with each trace's median and then apply a first-order Butterworth high-pass filter at 0.01 Hz. Filtered traces define

$$d_{ik} = \frac{\sqrt{T^{-1}\sum_t (a_i(t) - a_k(t))^2}}{\max_{u,v} \sqrt{T^{-1}\sum_t (a_u(t) - a_v(t))^2}} \in [0, 1]. \tag{16}$$

Validation uses $\alpha \in \{0.25, 0.5, 0.75\}$, $\rho \in \{0.1, 1, 10\}$, and $\varepsilon = 0.01$. The MM solver uses joint KL regularization, 10 BCD iterations, and 1,000 UOT iterations. Zebrafish uses RMS activity distances between 128-point traces from the evaluated 30-second windows, the same validation grid, and MM. Parameters are frozen before testing; one coupling supplies both directional rankings and Hungarian assignment.

E.5 Fixed-Encoder Matcher Controls

For the controls in Table 4, the NeuReL encoder and training atlas are frozen. The Sinkhorn control uses the frozen node-similarity initialization, unary temperature, and unmatched logits, with no relation-refinement updates. FGW uses the POT implementation (Flamary et al., 2021); FGW and unbalanced FGW (FUGW) both use ℓ_2-normalized node embeddings $\bar{\mathbf{z}}$ to define cross-population feature costs and within-population structural costs:

$$\begin{aligned} C^{\text{feat}}_{ij} &= \frac{\|\bar{\mathbf{z}}^Q_i - \bar{\mathbf{z}}^R_j\|_2^2}{4}, \\ C^Q_{ik} &= \frac{\|\bar{\mathbf{z}}^Q_i - \bar{\mathbf{z}}^Q_k\|_2^2}{4}, \qquad C^R_{jl} = \frac{\|\bar{\mathbf{z}}^R_j - \bar{\mathbf{z}}^R_l\|_2^2}{4}. \end{aligned} \tag{17}$$

Validation selects settings from Table 16 without encoder retraining, prioritizing real-candidate Top-1. Ties are broken in order by partial-assignment accuracy, MRR, Hungarian accuracy, and grid order. FUGW's retained-row-mass threshold is used for the auxiliary partial-assignment criterion during validation; it does not threshold the real-candidate scores reported in the main comparison. Candidate-score ties follow stable training-atlas slot order.

Table 16: Validation grids for fixed-encoder matcher controls. The FUGW solver uses the implementation and MM settings specified above.

Matcher	Settings
FGW	$\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$; at most 10,000 iterations; relative and absolute solver tolerances 10^{-9}.
FUGW	$\alpha \in \{0.25, 0.5, 0.75\}$, $\rho \in \{0.1, 1, 10\}$, $\varepsilon = 0.01$; retained-row-mass rejection threshold $\{0, 0.25, \ldots, 1.5\}$.