# Connectoformer: Predicting Synaptic Connectivity from Neuron Morphology

**Sidharth Goel**
Department of Computer Science
New York University
sg8603@nyu.edu

**Erdem Varol**
Department of Computer Science
New York University
ev2240@nyu.edu

## Abstract

Neuronal morphology constrains which neurons can form synaptic connections. Understanding how morphology shapes connectivity could help prioritize candidate synapses for proofreading and reveal structural features associated with circuit wiring. Recent connectomes provide the reconstructed morphologies and synapse annotations needed to build computational models that capture this relationship at scale. Training such models requires addressing the imbalance between the few neuron pairs that connect and the many that do not, as well as representing complex neuronal arbors as inputs to neural networks. Sampling nearby connected and unconnected pairs provides a more informative learning task, while the branching structure of neuronal arbors calls for a tokenization that preserves their geometry and topology. We introduce *Connectoformer*, a Transformer-based model that predicts directed connections from neuron morphology. *Connectoformer* decomposes each neuron skeleton into branch segments and represents each segment as a token containing its geometric features and position within the neuronal tree. A shared Transformer encodes the branch tokens of each neuron independently. To predict a connection, the model combines a score derived from whole-neuron representations with a score derived from cross-attention between the branches of the two neurons. The model is trained to predict connectivity and, using synapse location annotations as additional supervision, to identify branches likely to contain the neuron pair's synapses. On the FlyWire adult *Drosophila* connectome, *Connectoformer* predicts connected neurons from nearby unconnected neurons better than an arbor-overlap baseline. Compared with models that pool branch features into a single representation per neuron, branch-level interactions improve prediction most clearly for nearby and developmentally related candidates. We further evaluate generalization to held-out brain regions, developmental lineages, and neuron shapes, as well as to a second fly. Using synapse locations during training, *Connectoformer* also identifies branches likely to contain a neuron pair's synapses, providing a starting point for investigating where connections form along neuronal arbors. Code is available at `https://github.com/sgoe1/connectoformer`.

## 1 Introduction

Brains comprise neurons of diverse cell types, each extending a branching arbor through shared tissue. Neighboring arbors can interweave without connecting, while a localized contact between particular branches can contain many synapses [34, 40]. Understanding this selectivity requires relating neuronal shape to both partner selection and the distribution of synapses along the branches of each neuron.

Predicting partners and synaptic sites could guide reconstruction and connectivity experiments and, with longitudinal measurements, studies of structural plasticity [16]. Such predictions could also

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

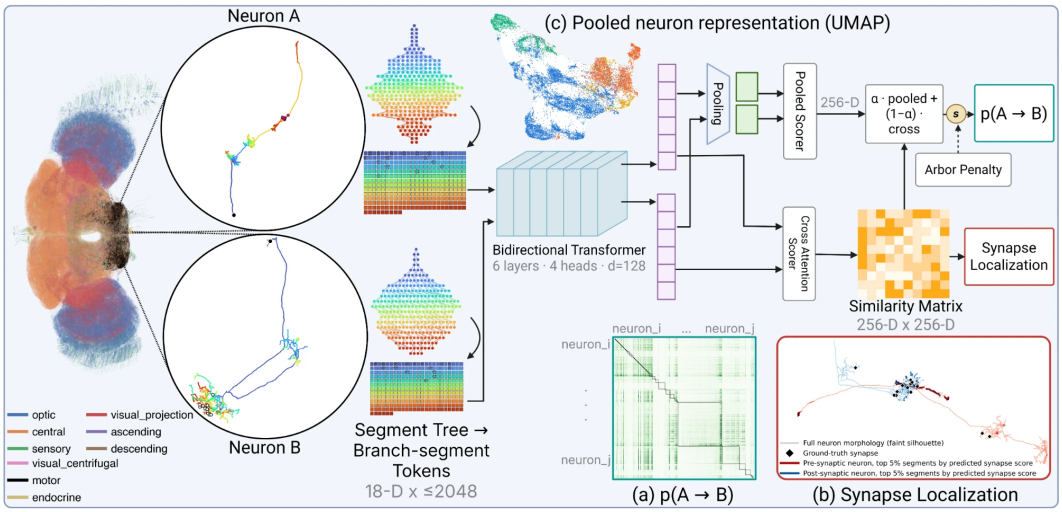



Figure 1: **Connectoformer.** *Left:* FlyWire neurons colored by super-class, with two example neurons (A, B). Each skeleton is split into branch segments ordered outward from the soma, and each neuron's segment tokens are encoded independently by a shared bidirectional Transformer. A pair is scored in two ways, by comparing pooled whole-neuron summaries and by letting each neuron's branches attend to its partner's branches, and the two scores are combined; an optional arbor penalty suppresses pairs whose arbors barely overlap. The branch-to-branch similarity matrix also ranks which branches likely carry the pair's synapses. **(a)** Predicted connectivity among held-out neurons grouped by neuroblast (black outlines; black dots mark true connections). **(b)** An example pair: segments with the top $5\%$ predicted synapse scores on the presynaptic (red) and postsynaptic (blue) neuron, with ground-truth synapses (diamonds). **(c)** UMAP of the pooled neuron representations of held-out neurons, colored by the same super-class colors as the brain at left; neurons group by region and cell identity (Appendix D.5).

inform computational models: in the *Drosophila* visual system, connectome-constrained models have already combined measured wiring with task-based learning to predict neural computation [23]. To explore this, we use static connectome reconstructions to ask how much partner and synaptic-location information the morphologies of two neurons contain.

Axons and dendrites must come close enough to form a synapse. Peters' rule relates expected connectivity to their spatial overlap [4, 30, 34], while the "potential-synapse" framework identifies sites where a synapse could form, whether or not one is present [48, 49]. Neural contact is shown to predict connectivity in *C. elegans* [4], and morphology-based models reproduce aspects of cortical organization in higher organism [35, 53]. Cable length models relate to branching and occupied volume of neurons [8]; across connectomes, Piazza and colleagues relate distributions of cable length, degree, and total synapse count to physical constraints and multiplicative growth [31]. These relationships describe contact opportunities and broad connectivity patterns, but is not specific enough to determine individual partners or synaptic sites [34, 40].

Morphology may contain further information about partner selection because neuronal shape develops alongside connectivity. Genetic programs regulate dendritic branching, and transcriptomic identity covaries with shape [12, 17]. Gene expression also predicts aspects of electrical or chemical connectivity in *C. elegans* [1, 21, 51]. During development, attractive and repulsive cues guide axons toward target regions, while branching, self-avoidance, and wiring costs constrain where neuronal processes can meet [7, 20, 27]. Recognition molecules contribute to selecting partners and synaptic sites among those contacts [5, 39, 40]. A reconstructed arbor is therefore a partial record of developmental and physical influences that also affect connectivity. Learning from its branches lets us test how much partner information remains in that record, without treating shape as a complete or causal explanation of synaptic specificity.

Electron-microscopy reconstructions now make this test possible by pairing neuronal morphologies with synapse annotations. These datasets span *C. elegans* and its development [57, 59], larval and adult *Drosophila* [10, 58], brain-wide volumes and reconstructed circuits in larval zebrafish [15, 50], and large mouse and human cortical volumes [29, 45].Specifically, we train and evaluate on

FlyWire's adult fruit-fly brain [10, 42] due to completeness of its annotation and large sample size of neurons and synapses. We also leverage a second adult fly [2] to ask whether learned morphology–connectivity relationships extends across individuals.

Methodologically, we introduce *Connectoformer* to predict directed connections in the FlyWire connectome as a function of neuron-pair morphologies. Furthermore, with synapse-location supervised labels during training, *Connectoformer* also ranks arbor branches likely to contain each pair's synapses.

To train *Connectoformer*, we convert each neuron skeleton into a variable-length sequence of non-branching segment tokens, ordered outward from the cell body and describing geometry, normalized position, and tree location. The variable-length sequence lets neurons with different numbers of branches share one encoder. This representation builds on morphometrics, topological descriptions, and learned tree and graph representations [3, 11, 19, 44, 56]. Whereas prior cross-attention groups points into neurons during reconstruction [52], we compare branches across two reconstructed neurons to retain pair-specific spatial information.

A shared bidirectional Transformer independently encodes each neuron using tree-position embeddings to retain each branch's place in the arbor. The model combines a score from pooled whole-neuron representations with a score from cross-attention between the candidate partners' arbor branches. Branch level cross-attention module also enables synapse localization predictions along the dendritic arbors. *Connectoformer* does not take as input cell-type labels, developmental annotations, and known connections of test neurons to emphasize learning morphological cues to drive predictions.

We compare *Connectoformer* with proximity, overlap, and lineage rules and MLPs using global summaries, pooled branch statistics, or both. The combined MLP matches the morphology inputs of *Connectoformer* + summary but pools individual branches. We test random and spatially contiguous neuron holdouts, unseen brain regions, entire neuron lineages, shape clusters, and generalizability across a new fly. Fixed cell-type pairs and dataset ablations further test whether predictions depend on neural identity or on specific particular model components.

Global shape embeddings improves discrimination among spatially plausible partners, while individual branches provide the clearest gains for nearby and same-lineage candidates. We also observe gains on held-out shapes and transfer to a second fly. No morphology-only architecture is clearly best in the unseen-region test, and *Connectoformer* does not improve over the matched MLP within fixed cell-type pairs. Synapse-location supervision supports above-chance presynaptic branch ranking, while postsynaptic localization remains weak. These results support ranking plausible partners and prioritizing presynaptic branches for further investigation.

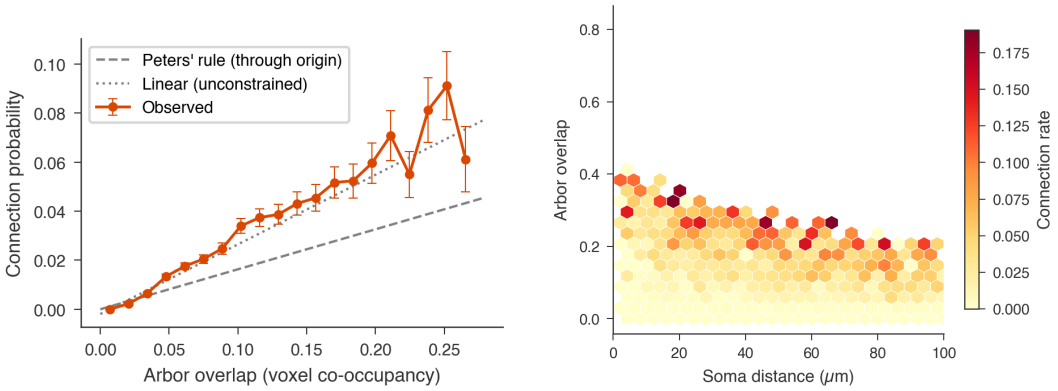



Figure 2: **Arbor overlap creates opportunity but does not select partners.** Ten million directed FlyWire pairs at $128^3$ voxelization. **Left:** empirical connection probability departs from both a through-origin Peters' rule fit (dashed) and an unconstrained linear fit (dotted). **Right:** connection rate as a function of soma distance and overlap. Among 12,270 pairs with soma distance below 30 $\mu$m and overlap above 0.05, only 2.4% connect. These plausible non-connections motivate the hard-negative task.

## 2 Related Work

**Geometry and connectivity.** Peters' rule and the potential-synapse framework model connectivity through axodendritic apposition [30, 48, 49]. Geometric overlap predicts cortical projection strength only up to a scale factor that varies with laminar and columnar location [46], and adjacency is strongly predictive in *C. elegans* [4]. In mouse visual cortex, similarity in feature tuning predicts connectivity beyond axodendritic proximity [9]. Generative models derive connectivity from appositions followed by pruning [35] or from dense overlap of pre- and postsynaptic sites [53]. Harth et al. [14] separate morphological constraints from cell-type, individual-cell, and subcellular specificity in dense reconstructions. We instead ask what a supervised skeleton model predicts beyond geometric opportunity.

**Genetic and developmental models.** Complementary approaches predict connectivity from gene expression and spatial information, including functional connectivity between brain regions [33] and gap-junction wiring in *C. elegans* [21], or model its development through cell type, birth time, soma distance, reciprocity, and pruning [36]. Our morphology-only task measures predictive information available without molecular or developmental annotations at inference.

**Learned neuron representations.** Morphology has been represented using hand-crafted morphometrics [44], topological descriptors [19], NBLAST similarity [6], tree encoders [3], graph attention [56], and self-supervised or joint morphology-connectome models [11, 24]. Point Affinity Transformer [52] also constructs cross-attention affinities, but for within-volume point grouping rather than connectivity between independently held-out neurons. NCPNet [55] predicts links from partial connectivity, whereas NTAC [43] assigns cell types from connectivity. Our inductive task instead withholds test-neuron graph edges.

## 3 Data, Representation, and Evaluation

**FlyWire connectome dataset.** We use $138{,}584$ neurons and $3{,}732{,}460$ directed connections from FlyWire release `v783` (Appendix A), at the standard threshold of at least five synapses [10]. Given two SWC skeletons $(a, b)$, the model predicts whether $a \to b$. Cell type, lineage, neurotransmitter, neuron identifier, and graph features are withheld from the primary morphology predictors. Annotations are used only to define splits and evaluation subsets, and as inputs to separate identity-only baselines. We evaluate the frozen FlyWire models on the male adult CNS connectome [2] to test cross-individual transfer.

**From SWC trees to tokens.** Figure 3 illustrates how each soma-rooted SWC becomes a breadth-first sequence of 18-D branch-segment tokens describing local geometry, tree context, and soma-centered, extent-normalized position. Learned depth and sibling-rank embeddings encode each segment's place in the tree; up to $16$ deterministic subtree-summary tokens add multiscale context. Appendix B provides the complete feature definitions and tensor shapes.

Neurons exceeding $2{,}048$ branch tokens ($5{,}674$; $4.1\%$) are excluded, not truncated; retained neurons are encoded in full. The pair scorer then uses up to $K = 256$ approximately evenly spaced encoded tokens per neuron. This selection limits pairwise computation, not the encoder's context: unselected branches can still influence the retained representations.

**Global summary and overlap.** The 25-D whole-neuron summary describes global shape, including six Sholl counts (how often the arbor crosses spheres of increasing radius around the soma), which no single segment token captures. The matched MLP concatenates it with $54$ token statistics (mean, maximum, and standard deviation of each token feature), giving a 79-D vector; *Connectoformer* + summary receives the same summary alongside the individual tokens. Arbor overlap is a separate scalar computed from voxelized co-occupancy in the shared anatomical frame. The fixed arbor prior adds a one-sided log penalty that suppresses pairs with near-zero overlap and leaves the rest unchanged; a variant that learns how to use overlap is reported as a control in Appendix D.

**Splits and uncertainty.** Table 1 summarizes the three splits; no neuron appears in more than one partition. **Random** holds out randomly chosen neurons, testing prediction for new neurons from familiar populations; five independent four-fold partitions give 20 held-out folds. **Region** tests transfer to an anatomically distinct area: it trains on central and other superclasses and tests on the optic lobe. Its checkpoints are selected on held-out visual projection and centrifugal neurons ("bridge"

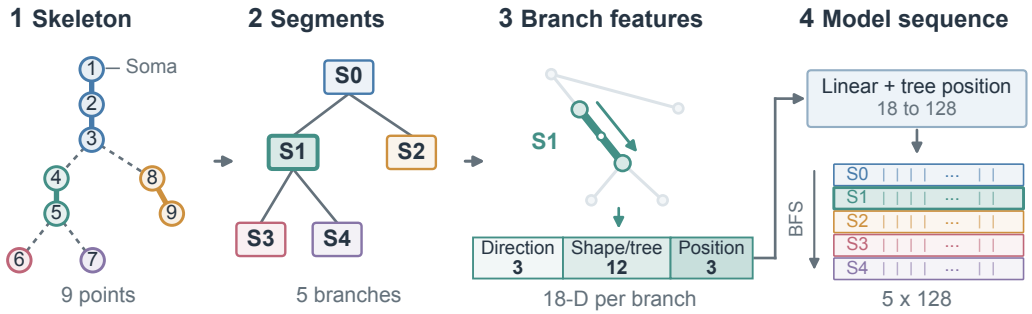



Figure 3: **From skeletons to topology-aware tokens.** (1) An SWC skeleton records point coordinates, radii, and parent links. (2) Non-branching point sequences form nodes in a rooted segment tree; fork points belong to the upstream segment, and single-point terminal segments are retained. Colors track the same segments across panels. (3) Branch S1 illustrates the 18-D feature vector: three parent-relative direction components, twelve geometry/tree statistics, and three soma-centered, extent-normalized midpoint coordinates. The arrow and white marker show direction and midpoint. (4) Breadth-first (BFS) ordering yields a sequence of branch vectors, projected to 128 dimensions and augmented with learned depth and sibling-rank embeddings before bidirectional encoding. Bars are symbolic, not measured feature values. The example omits the additional subtree-summary tokens and masked batch padding; BFS ordering does not impose a causal attention mask.

neurons that link the two regions), and its five seeds vary optimization. **Patch** trains on neurons whose somas lie within 75 $\mu$m of 30 random centers and tests on the rest, so held-out neurons form contiguous regions rather than being interleaved with training neurons; its five seeds vary both the partition and optimization. Group transfer (Section 5.3) holds out entire lineages or shape clusters in five folds with one seed each. Unless a caption states otherwise, results are mean $\pm$ SE over five seeds, averaging the four Random folds within each seed first. Paired comparisons use the same units, two-sided $t$ intervals, and Holm correction across contrasts (Appendix C).

**Metrics.** Overall AUROC and average precision (AP) compare true connections against negatives that are half random pairs and half nearby non-connected pairs. **Hard** AUROC uses only nearby negatives: non-connected neurons whose somas are among the source neuron's nearest neighbors. **In-Type** negatives are nearby non-connected neurons that also share the source's neuroblast group (neurons born from the same neural stem cell); despite the name, it does not require the same cell type. These subsets stop distance and developmental lineage from serving as shortcuts. Sampling details and metrics at the true population connection rate are in Appendix C.

## 4 Model

Four kinds of signal could indicate that neuron $a$ connects to neuron $b$ at branches $i$ and $j$: the neurons' overall shapes, the compatibility of the two branches, their spatial proximity, and where the branches sit in each arbor. We use the following decomposition to formulate this relation:

$$\begin{aligned}\mathcal{S}(a,b,i,j) = C &+ \mathcal{S}_{\text{global}}(c_a, c_b) + \mathcal{S}_{\text{local}}(h_{ai}, h_{bj}) \\ &+ \mathcal{S}_{\text{distance}}(r_{ai}, r_{bj}, d_{ij}) + \mathcal{S}_{\text{tree}}(i, j, T_a, T_b).\end{aligned} \tag{1}$$

Table 1: **Dataset and split statistics.** Neurons and directed positive pairs ($\geq$ 5 synapses), in thousands, for each partition of the 138,584 FlyWire neurons; pairs spanning two partitions are discarded. Random: mean over 20 folds; Patch: mean $\pm$ SD over five seeds; Region: one fixed partition. Counts precede the 2,048-token filter (Section 3).

| | Neurons (k) | | | Positive pairs (k) | | | |
|---|---|---|---|---|---|---|---|
| **Split** | Train | Val | Test | Train | Val | Test | Cross |
| Random | 93.5 | 10.4 | 34.6 | 1,702 | 21 | 233 | 1,777 |
| Region | 52.6 | 8.2 | 77.8 | 1,431 | 78 | 1,532 | 691 |
| Patch | $95.6 \pm 4.1$ | $10.6 \pm 0.5$ | $32.4 \pm 4.5$ | $1{,}862 \pm 171$ | $23 \pm 3$ | $436 \pm 90$ | $1{,}411 \pm 103$ |

Here $C$ is a constant offset, $c$ summarizes a neuron, $h$ represents a branch, $r$ and $d$ describe branch positions and their separation, and $T$ is the rooted arbor topology. This is an organizing framework, not four separately optimized loss terms or a constraint that the implemented score be additive. Global and local evidence enter the two scoring paths, tree context enters the token encoder, and spatial feasibility enters through geometry and optional arbor overlap (illustrated in Appendix Figure 7).

**Baselines.** Three rule-based controls score pairs by shared hemilineage (a developmental subdivision of a neuroblast lineage), soma proximity, and voxelized arbor overlap. We compare bilinear-scoring MLPs on the 25-D global summary (*summary MLP*), the 54-D token statistics (*token-statistics MLP*), or their 79-D concatenation (*matched MLP*), separating whole-neuron shape from averaged branch features. The matched MLP sees the same features as *Connectoformer* + summary, but averaging discards the individual branches that cross-attention compares; it matches input content, not parameter count.

**Independent bidirectional encoder.** Each neuron is encoded independently by a shared six-layer, four-head Transformer [54] ($d = 128$, FFN $= 256$). Self-attention is bidirectional because BFS order is a serialization of a tree, not a temporal or autoregressive direction. Neurons do not interact in this trunk. Soma-centered coordinates remove translation and extent normalization removes scale from the midpoint channels. Orientation information remains frame-dependent; rotation invariance is not imposed. Coordinate sensitivity is evaluated in Appendix D.4.

**Pooled and cross-attention scoring.** The pooled path compares whole neurons; the cross-attention path compares branches. The pooled path uses four learned attention queries over all final token states to form four static summary slots per neuron. Separate source/target projections score all 16 slot pairs bilinearly, followed by log-mean-exp aggregation (a smooth maximum). The slots are independent of the partner and can be cached. The cross-attention path depends on the partner: tokens from $a$ query tokens from $b$, then tokens from $b$ query the updated source states. Separate normalized 16-D pre/post projections form a $K_a \times K_b$ compatibility matrix, scaled by a learned positive temperature. Log-mean-exp over valid entries produces the local logit. Writing these logits as $s_{\text{pool}}$ and $s_{\text{cross}}$, the implemented morphology score is

$$s_{\text{morph}}(a, b) = \alpha s_{\text{pool}}(a, b) + (1 - \alpha) s_{\text{cross}}(a, b), \qquad \alpha = \sigma(\alpha_{\text{logit}}). \tag{2}$$

The scalar $\alpha_{\text{logit}}$ is learned jointly with the network, initialized to give equal weight to both paths. Thus broad compatibility and branch-to-branch interactions contribute to the same decision. Cross-attention costs $O(K^2)$ per candidate pair. Independent neuron encoding is performed once per checkpoint and cached, but the partner-specific scoring step is repeated for each pair.

**Training.** We train with binary cross-entropy on balanced positive/negative batches whose negatives are half uniform and half soma-neighborhood samples. In dense-subset training, each step encodes a densely interconnected subset of up to 10,240 neurons and scores many pairs within it; reusing each encoding across pairs makes training efficient and gives every architecture the same number of training pairs. The main *Connectoformer* model also uses two auxiliary losses built from training-set synapse locations, which are never model inputs at inference. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{conn}} + 0.02\mathcal{L}_{\text{node}} + 0.1\mathcal{L}_{\text{rank}}. \tag{3}$$

The node term trains branch scores to match where synapses are observed, and the ranking term scores synapse-bearing branches above other branches. These are training objectives, distinct from the four sources of evidence in Eq. 1; Section 5.4 tests their effect. The Transformer is optimized with AdamW at learning rate $3 \times 10^{-4}$, weight decay $10^{-4}$, a 20-epoch warmup, and at most 1,000 epochs with validation-based early stopping. Complete optimization and checkpoint-selection details, including baseline optimization and auxiliary weights, are in Appendix B.

# 5 Results

## 5.1 Morphology predicts beyond spatial opportunity

Table 2 compares rules, whole-neuron summaries, branch-token models, and overlap-augmented models on overall and hard-case discrimination.

Simple rules explain opportunity more readily than partner choice. Arbor overlap is competitive overall, consistent with the geometric opportunity in Fig. 2, but its performance drops on nearby

Table 2: **Connectivity prediction across three held-out splits.** AUROC mean with SE as subscript over five seeds (Appendix Table 6). MLP inputs: 54-D token statistics, 25-D summary, or both (79-D). **Bold**: highest value among the rule and morphology-only rows, and separately among the arbor-prior rows. Paired effects: Supp. Table 8; learned overlap: Supp. Table 10.

| | Random | | | Region | | | Patch | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **Overall** | **Hard** | **In-Type** | **Overall** | **Hard** | **In-Type** | **Overall** | **Hard** | **In-Type** |
| *Rules* | | | | | | | | | |
| Hemilineage | $.519_{.001}$ | $.434_{.002}$ | $.171_{.002}$ | $.537_{.000}$ | $.488_{.000}$ | $.242_{.000}$ | $.517_{.014}$ | $.404_{.019}$ | $.157_{.018}$ |
| Soma proximity | $.479_{.000}$ | $.050_{.001}$ | $.047_{.001}$ | $.483_{.000}$ | $.056_{.000}$ | $.057_{.000}$ | $.492_{.007}$ | $.090_{.008}$ | $.087_{.008}$ |
| Arbor overlap | $.841_{.001}$ | $.686_{.002}$ | $.569_{.002}$ | $\mathbf{.834_{.000}}$ | $.668_{.000}$ | $.605_{.000}$ | $\mathbf{.864_{.003}}$ | $.728_{.006}$ | $.595_{.017}$ |
| *Morphology only* | | | | | | | | | |
| MLP (54-D tokens) | $.787_{.002}$ | $.767_{.003}$ | $.759_{.003}$ | $.671_{.003}$ | $.659_{.003}$ | $.657_{.003}$ | $.765_{.007}$ | $.733_{.008}$ | $.710_{.009}$ |
| MLP (25-D summary) | $.839_{.001}$ | $.788_{.002}$ | $.770_{.002}$ | $.692_{.002}$ | $.689_{.003}$ | $.681_{.005}$ | $.812_{.005}$ | $.748_{.006}$ | $.713_{.006}$ |
| MLP (79-D combined) | $.841_{.002}$ | $.789_{.003}$ | $.770_{.003}$ | $.694_{.010}$ | $.692_{.009}$ | $.688_{.013}$ | $.809_{.005}$ | $.743_{.005}$ | $.705_{.005}$ |
| *Connectoformer* (tokens) | $.823_{.001}$ | $.812_{.002}$ | $.809_{.002}$ | $.702_{.003}$ | $\mathbf{.701_{.004}}$ | $\mathbf{.700_{.004}}$ | $.798_{.005}$ | $.779_{.007}$ | $.761_{.006}$ |
| *Connectoformer* + summary | $\mathbf{.862_{.001}}$ | $\mathbf{.842_{.001}}$ | $\mathbf{.834_{.002}}$ | $.677_{.010}$ | $.673_{.013}$ | $.663_{.015}$ | $.824_{.003}$ | $\mathbf{.790_{.006}}$ | $\mathbf{.769_{.004}}$ |
| *Morphology + fixed arbor prior* | | | | | | | | | |
| MLP (25-D summary) | $.909_{.001}$ | $.823_{.002}$ | $.804_{.002}$ | $.852_{.011}$ | $.709_{.023}$ | $.706_{.026}$ | $.909_{.005}$ | $.825_{.011}$ | $.785_{.010}$ |
| MLP (79-D combined) | $.911_{.001}$ | $.825_{.002}$ | $.806_{.001}$ | $.850_{.005}$ | $.706_{.011}$ | $.704_{.011}$ | $.911_{.005}$ | $.827_{.011}$ | $.787_{.011}$ |
| *Connectoformer* + summary | $\mathbf{.916_{.001}}$ | $\mathbf{.837_{.002}}$ | $\mathbf{.826_{.002}}$ | $\mathbf{.867_{.003}}$ | $\mathbf{.745_{.004}}$ | $\mathbf{.751_{.005}}$ | $\mathbf{.913_{.005}}$ | $\mathbf{.833_{.011}}$ | $\mathbf{.802_{.010}}$ |

and same-lineage negatives. Global morphology retains considerably more discrimination on these strata: on Random Hard, for example, the summary MLP reaches 0.788 AUROC versus 0.686 for overlap. Soma proximity and hemilineage fall far below chance on the strata built from their own signal: Hard negatives are the source neuron's soma neighbors and In-Type negatives share its neuroblast group, so each rule ranks those negatives above true partners.

Overlap remains useful information. Adding the fixed arbor prior raises overall AUROC for every scorer, and with it *Connectoformer* is the best model on all three splits. Against the matched MLP with the same prior, its hard-case advantage is consistent: it holds in every seed on all three splits, and is significant after Holm correction on Random and for Patch In-Type (Table 8). The prior acts as a feasibility filter rather than a uniform improvement: on Random it slightly lowers *Connectoformer*'s Hard and In-Type AUROC. Learned-overlap controls are in Appendix D. Figure 4 shows how the overlap-aware model ranks one neuron's candidate partners among all held-out neurons.

### 5.2 Branch structure adds hard-case discrimination

The morphology-only rows reveal a useful tradeoff. Tokens-only *Connectoformer* trails the summary MLP overall on Random, yet exceeds all three MLPs on Hard and In-Type. On Random and Patch, adding the global summary improves the token model, suggesting that whole-neuron shape and branch-level representations are complementary. Against the matched MLP, *Connectoformer* + summary improves Hard and In-Type AUROC by five to seven points on Random (leading in all 20 folds) and by similar margins on Patch, and improves overall AUROC on Random by two points; all of these gains are significant after Holm correction (Table 8). On Patch without the prior, arbor overlap is strongest overall, whereas *Connectoformer* + summary leads on Hard and In-Type. The token model's advantage lies in telling apart difficult candidate pairs, not in improving every ranking task.

The held-out region changes the ordering. There, tokens-only *Connectoformer* has the highest morphology-only Hard and In-Type AUROC. It clearly beats the token-statistics MLP, which sees the same features averaged, and edges out the summary MLP in every seed, but is statistically indistinguishable from the matched MLP. Adding the global summary does not help: *Connectoformer* + summary falls slightly below the matched MLP, within seed-to-seed variation. Branch-level structure therefore transfers to an unseen region better than the same features averaged, but no morphology model is clearly best there.

**Which components matter?** Removing the global summary lowers Random overall AUROC by about four points and Hard AUROC by three (Table 2). Removing dense-subset training leaves overall AUROC unchanged but lowers Hard AUROC by about five points (Appendix D.3). Tree-position

embeddings make a smaller contribution. These gains characterize the model and its training recipe together, not cross-attention alone. Inference-time feature masking, which measures reliance rather than retrained information, is also in Appendix D.3.

**Does morphology merely recover cell identity?** An MLP given only annotated identity labels underperforms the summary MLP, and adding these labels to morphology changes little. A stricter test fixes both the source and target cell types and asks which of those pairs connect, across 128 such type pairs. Both models do slightly better than chance (AUROC averaged over type pairs about 0.56), but *Connectoformer* shows no gain over the matched MLP (Appendix D.4). Morphology therefore captures more than an identity lookup, but the architectural advantage does not extend to choosing partners within a single cell type.

Figure 5 shows held-out, neuroblast-sorted score matrices for a qualitative view of these comparisons.

**Where does the hard-case advantage appear?** Figure 6 maps where *Connectoformer*'s Hard advantage over the matched MLP arises, using six Random test folds from a separate evaluation panel. The maps describe held-out pair ranking, not per-neuron accuracy or unseen-region transfer (Appendix D.6).

### 5.3 Generalization beyond random neuron splits

Holding out entire developmental lineages or shape clusters tests whether prediction extends to unfamiliar kinds of neurons, not just new neurons from familiar groups (Table 3). Lineage folds hold out annotated hemilineages. Shape clusters are 32 clusters of standardized 25-D shape summaries, built without connectivity labels; whole clusters are assigned to folds, so test shapes are absent from training. The overlap-augmented rows here use the learned-overlap variant (Appendix D) rather than the fixed prior, because it lets each model learn how much to rely on overlap instead of imposing the same fixed filter on both.

The clearest gains are on unfamiliar shapes and same-lineage discrimination. *Connectoformer* has higher In-Type AUROC in all four conditions, with paired intervals excluding zero for both shape-cluster conditions and for lineage with learned overlap; with learned overlap, held-out shape clusters also show gains on Hard and overall AUROC. Elsewhere, overall differences are near zero or favor the MLP, and lineage Hard gains are not distinguishable from zero (intervals in Appendix C).

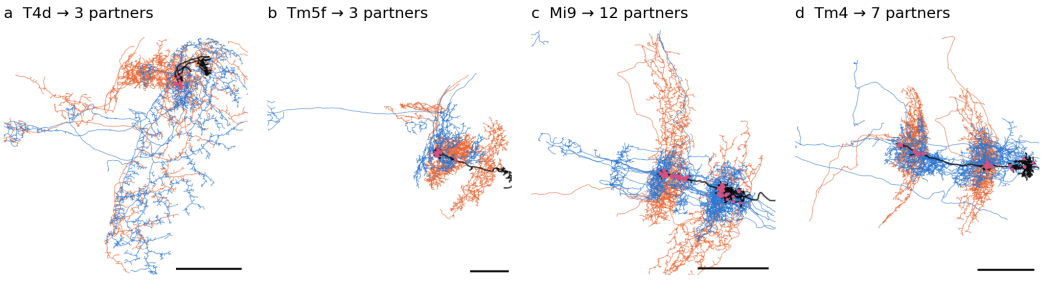



Figure 4: **Ranked partners for four held-out presynaptic neurons.** *Connectoformer* + summary + learned overlap (Random seed 42, fold 0) ranks all 33,233 held-out neurons as candidate partners of each source. **Top:** frontal projection of the source (black), its held-out postsynaptic partners (blue), its highest-ranked non-partners (orange), and its presynaptic sites onto those partners (magenta); scale bars 50 $\mu$m. **Bottom:** the model's top five candidates (blue: connected; orange: not connected) with each candidate's rank under raw arbor overlap. In (a, b) the model ranks a partner first that overlap ranks far lower; in (c, d) overlap ranks a partner first while the model prefers large wide-field neurons. Top-ten lists are in Appendix Figure 8.

**Transfer to another fly.** On the male *Drosophila* CNS, models trained on FlyWire and applied without retraining reach Hard AUROC 0.739 for *Connectoformer*, 0.701 for the matched MLP, and 0.623 for overlap on the same pairs. The hard-case advantage therefore carries over to a second fly without retraining, although the MLP is slightly better on easy negatives (Appendix F.1).

### 5.4 Localization is a coarse auxiliary signal

Cross-attention produces pre- and postsynaptic branch scores, a spatial output that an averaged-feature MLP cannot supply. We test whether these scores rank the branches that actually carry a pair's synapses above the neuron's other branches. Hit@$k$ counts a success when a true synaptic branch is among the $k$ highest-scoring branches on that side of the pair. On 1,000 held-out connected pairs, a true presynaptic branch is in the top ten for 41.0% of pairs, three times the chance rate of 13.6%; the postsynaptic signal is weak (14.4% versus 11.1%; Appendix Table 13). The scores therefore prioritize branches, especially presynaptically, rather than locate synapses precisely.

This signal comes from the auxiliary objective: trained without the synapse-location losses (Eq. 3), branch scores rank synaptic branches at chance, and adding them costs less than one point of connectivity AUROC. Appendix E gives the full retrieval metrics and projection-error controls.

## 6 Discussion

Arbor overlap identifies spatial opportunity, while whole-neuron shape improves partner discrimination. Retaining individual branches adds most when candidates are already nearby, suggesting that their arrangement helps distinguish plausible partners. This advantage does not hold uniformly: in the unseen-region test, the best morphology models perform similarly.

**Biological implications.** The results also indicate where shape carries partner information and where it runs out. Morphology separates a neuron's partners from nearby non-partners born from the same neural stem cell, which a lineage rule ranks above true partners, and branch-level structure helps most on these candidates. Branch layout therefore carries partner information beyond lineage

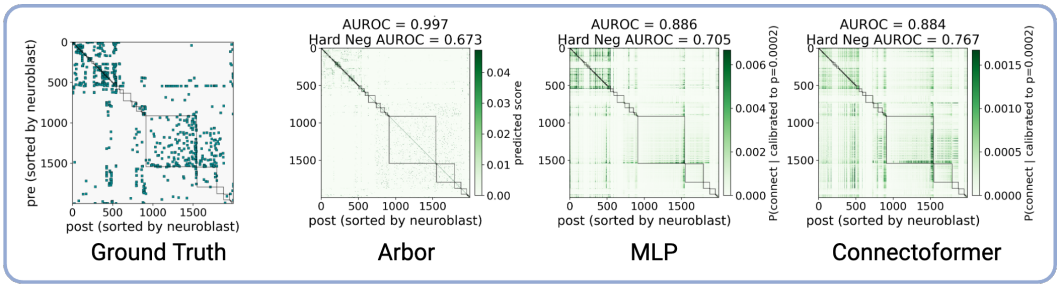



Figure 5: **Predicted connectivity matrices on a held-out neuroblast-sorted test subset.** Ground truth (left) shows synaptic connections as teal points; the remaining panels show scores for arbor overlap, the MLP, and *Connectoformer* on the same neurons, with neuroblast blocks outlined. Learned-scorer logits are rescaled to the true population connection rate by prior shift [38]. The displayed metrics describe this visualization subset; repeated-run comparisons are in Tables 2 and 8.

Table 3: **Five-fold group-disjoint transfer.** AUROC mean ± SD across the five held-out lineage or shape-cluster groups. Each fold uses the same optimization seed, so the displayed SD quantifies held-out-group variability. Full paired AUROC/AP intervals are in Appendix C.

| Input | Held out | Method | Overall | Hard | In-Type |
|---|---|---|---|---|---|
| Morphology | Lineage | Matched MLP | .796±.055 | .733±.083 | .702±.064 |
| | | *Connectoformer* | .787±.035 | .779±.047 | .737±.039 |
| | Shape | Matched MLP | .827±.060 | .770±.106 | .741±.080 |
| | | *Connectoformer* | .827±.054 | .809±.077 | .780±.064 |
| + learned overlap | Lineage | Matched MLP | .962±.004 | .827±.021 | .759±.038 |
| | | *Connectoformer* | .956±.007 | .849±.034 | .805±.039 |
| | Shape | Matched MLP | .938±.019 | .857±.041 | .787±.022 |
| | | *Connectoformer* | .948±.023 | .888±.051 | .851±.047 |

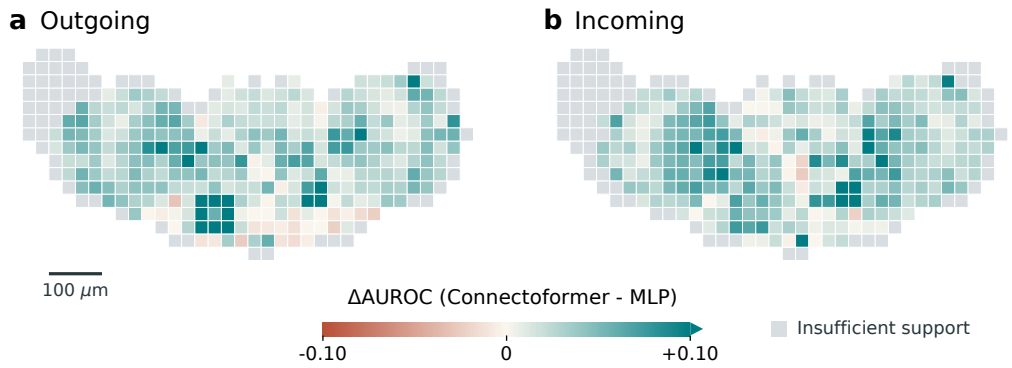



Figure 6: **Anatomical distribution of the hard-case ranking advantage.** Directed pairs are grouped by the **(a)** source soma (outgoing) or **(b)** target soma (incoming) in $25 \times 25\,\mu\mathrm{m}$ XY bins, pooling all depths. Color is *Connectoformer* minus matched 79-D MLP AUROC on identical positive and hard-negative pairs: teal favors *Connectoformer*, orange favors the MLP, and cream denotes near-zero difference. **Gray is insufficient support, not zero gain**: fewer than 30 unique pairs of either class in at least one of six held-out evaluations. Gray bins may still contain evaluated pairs. Faint points show all somas; performance uses pairs with both endpoints held out. Both models use morphology and the global summary, without overlap. Fold differences are averaged within seed, then equally across five seeds.

identity, consistent with developmental programs shaping where branches go. Shape runs out within fixed cell-type pairs, where morphology retains only modest signal, and on the postsynaptic side, where branch scores are barely above chance. One possibility is that recognition molecules such as Dpr–DIP act at contact sites without substantially reshaping the arbor [5, 40]. Comparing prediction errors with recognition-molecule expression would test this explanation.

**Practical use.** Because *Connectoformer*'s advantage lies in choosing among nearby candidates, its natural use is ranking a neuron's plausible partners, for example to prioritize which connections to test with trans-synaptic labeling or paired recordings, rather than screening all pairs. Because the models need only skeletons, they can also propose connections where neuron shapes are known but synapses are not, such as single-neuron light-microscopy reconstructions registered to a common template [28] or X-ray nano-tomography volumes, where synapse analysis has relied on correlative electron microscopy [22]; transfer to the male CNS without retraining supports this within *Drosophila*. Presynaptic branch scores can further narrow where to look for a pair's synapses. These uses are proposed rather than evaluated end to end, and absolute precision at the population connection rate is low, so predictions should guide experiments and review rather than replace synapse reconstruction.

**Scope and next steps.** The findings support predictive morphological regularities within adult *Drosophila*, including transfer to unseen shape clusters and another fly. Cross-species transfer remains unestablished: repeated *C. elegans* probes show no robust benefit from FlyWire pretraining over raw morphology (Appendix F.2). Stronger cross-species principles would require repeated gains over raw morphology and geometry across harmonized, matched-scale connectomes with within-species controls. Because synapses form at specific subcellular sites within a partner's arbor [39], more spatially resolved supervision and compartmental, molecular, or ultrastructural inputs are natural directions for improving synapse placement.

**Limitations.** Training excludes the $4.1\%$ of neurons that are above branch $2{,}048$ tokens; the localization panel includes them and reports them separately. Matching morphology inputs does not match parameter count or the Transformer's auxiliary supervision, so comparisons evaluate complete model recipes, not cross-attention in isolation. Region checkpoints are selected on bridge neurons, which are mostly long-range projection neurons, whereas the optic-lobe test set is dominated by compact local types, so this proxy may select imperfectly. Intervals for the exact-type and male-CNS panels hold one trained model fixed and may understate uncertainty when pairs share neurons. Molecular recognition and compartment identity are not supplied to the model.

## Acknowledgments and Disclosure of Funding

This work was supported by NIH grant R00MH128772 and by Google Cloud Platform. We thank NYU IT High Performance Computing for providing the computational resources, services, and staff expertise that enabled the experiments in this work. We also thank Alexander Ratzan for feedback on the writing.

## References

[1] Dániel L. Barabási and Albert-László Barabási. A genetic model of the connectome. *Neuron*, 105(3):435–445.e5, 2020. doi: 10.1016/j.neuron.2019.10.031.

[2] Stuart Berg, Isabella R. Beckett, Marta Costa, Philipp Schlegel, et al. Sexual dimorphism in the complete Drosophila male central nervous system connectome. *Cell*, 189(18):5504–5526.e15, 2026. doi: 10.1016/j.cell.2026.08.015.

[3] Hanbo Chen, Jiawei Yang, Daniel Maxim Iascone, Lijuan Liu, Lei He, Hanchuan Peng, and Jianhua Yao. TreeMoCo: contrastive neuron morphology representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 25060–25073, 2022.

[4] Steven J. Cook, Cristine A. Kalinski, and Oliver Hobert. Neuronal contact predicts connectivity in the C. elegans brain. *Current Biology*, 33(11):2315–2320.e2, 2023.

[5] Filip Cosmanescu, Phinikoula S. Katsamba, Alina P. Sergeeva, Goran Ahlsen, Saurabh D. Patel, Joshua J. Brewer, Liming Tan, Shuwa Xu, Qi Xiao, Sonal Nagarkar-Jaiswal, Aljoscha Nern, Hugo J. Bellen, S. Lawrence Zipursky, Barry Honig, and Lawrence Shapiro. Neuron-subtype-specific expression, interaction affinities, and specificity determinants of DIP/Dpr cell recognition proteins. *Neuron*, 100(6):1385–1400.e6, 2018.

[6] Marta Costa, James D. Manton, Aaron D. Ostrovsky, Steffen Prohaska, and Gregory S. X. E. Jefferis. NBLAST: rapid, sensitive comparison of neuronal structure and construction of neuron family databases. *Neuron*, 91(2):293–311, 2016.

[7] Hermann Cuntz, Friedrich Forstner, Alexander Borst, and Michael Häusser. One rule to grow them all: a general theory of neuronal branching and its practical application. *PLoS Computational Biology*, 6(8):e1000877, 2010. doi: 10.1371/journal.pcbi.1000877.

[8] Hermann Cuntz, Alexandre Mathy, and Michael Häusser. A scaling law derived from optimal dendritic wiring. *Proceedings of the National Academy of Sciences*, 109(27):11014–11018, 2012. doi: 10.1073/pnas.1200430109. URL `https://doi.org/10.1073/pnas.1200430109`.

[9] Zhuokun Ding, Paul G. Fahey, Stelios Papadopoulos, Eric Y. Wang, et al. Functional connectomics reveals general wiring rule in mouse visual cortex. *Nature*, 640(8058):459–469, 2025.

[10] Sven Dorkenwald, Arie Matsliah, Amy R. Sterling, et al. Neuronal wiring diagram of an adult brain. *Nature*, 634(8032):124–138, 2024.

[11] Yimin Fan, Yaxuan Li, Yunhua Zhong, Liang Hong, Lei Li, and Yu Li. Learning meaningful representation of single-neuron morphology via large-scale pre-training. *Bioinformatics*, 40 (Supplement 2):ii128–ii136, 2024. doi: 10.1093/bioinformatics/btae395.

[12] Nathan W. Gouwens, Staci A. Sorensen, Fahimeh Baftizadeh, Agata Budzillo, Brian R. Lee, et al. Integrated morphoelectric and transcriptomic classification of cortical GABAergic cells. *Cell*, 183(4):935–953.e19, 2020. doi: 10.1016/j.cell.2020.09.057.

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017.

[14] Philipp Harth, Daniel Udvary, Jan Boelts, Jakob H. Macke, Daniel Baum, Hans-Christian Hege, and Marcel Oberlaender. Dissecting origins of wiring specificity in dense reconstructions of neural tissue. *bioRxiv*, 2025. doi: 10.1101/2024.12.14.628490.

[15] David Grant Colburn Hildebrand, Marcelo Cicconet, Russel Miguel Torres, Woohyuk Choi, Tran Minh Quan, Jungmin Moon, Arthur Willis Wetzel, Andrew Scott Champion, Brett Jesse Graham, Owen Randlett, George Scott Plummer, Ruben Portugues, Isaac Henry Bianco, Stephan Saalfeld, Alexander David Baden, Kunal Lillaney, Randal Burns, Joshua Tzvi Vogelstein, Alexander Franz Schier, Wei-Chung Allen Lee, Won-Ki Jeong, Jeff William Lichtman, and Florian Engert. Whole-brain serial-section electron microscopy in larval zebrafish. *Nature*, 545(7654):345–349, 2017. doi: 10.1038/nature22356. URL `https://doi.org/10.1038/nature22356`.

[16] Anthony Holtmaat and Karel Svoboda. Experience-dependent structural synaptic plasticity in the mammalian brain. *Nature Reviews Neuroscience*, 10(9):647–658, 2009. doi: 10.1038/nrn2699. URL `https://doi.org/10.1038/nrn2699`.

[17] Shiho Jinushi-Nakao, Ramanathan Arvind, Reiko Amikura, Emi Kinameri, Andrew Winston Liu, and Adrian Walton Moore. Knot/Collier and Cut control different aspects of dendrite cytoskeleton and synergize to define final arbor shape. *Neuron*, 56(6):963–978, 2007. doi: 10.1016/j.neuron.2007.10.031.

[18] John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[19] Lida Kanari, Paweł Dłotko, Martina Scolamiero, Ran Levi, Julian Shillcock, Kathryn Hess, and Henry Markram. A topological representation of branching neuronal morphologies. *Neuroinformatics*, 16(1):3–13, 2018.

[20] Alex L. Kolodkin and Marc Tessier-Lavigne. Mechanisms and molecules of neuronal wiring: a primer. *Cold Spring Harbor Perspectives in Biology*, 3(6):a001727, 2011. doi: 10.1101/cshperspect.a001727.

[21] István A. Kovács, Dániel L. Barabási, and Albert-László Barabási. Uncovering the genetic blueprint of the C. elegans nervous system. *Proceedings of the National Academy of Sciences*, 117(52):33570–33577, 2020. doi: 10.1073/pnas.2009093117.

[22] Aaron T. Kuan, Jasper S. Phelps, Logan A. Thomas, Tri M. Nguyen, Julie Han, Chiao-Lin Chen, Anthony W. Azevedo, John C. Tuthill, Jan Funke, Peter Cloetens, Alexandra Pacureanu, and Wei-Chung Allen Lee. Dense neuronal reconstruction through X-ray holographic nano-tomography. *Nature Neuroscience*, 23(12):1637–1643, 2020. doi: 10.1038/s41593-020-0704-9.

[23] Janne K. Lappalainen, Fabian D. Tschopp, Sridhama Prakhya, Mason McGill, Aljoscha Nern, Kazunori Shinomiya, Shin-ya Takemura, Eyal Gruntman, Jakob H. Macke, and Srinivas C. Turaga. Connectome-constrained networks predict neural activity across the fly visual system. *Nature*, 634(8036):1132–1140, 2024. doi: 10.1038/s41586-024-07939-3. URL `https://doi.org/10.1038/s41586-024-07939-3`.

[24] Minghui Liao, Guojia Wan, and Bo Du. Joint learning neuronal skeleton and brain circuit topology with permutation invariant encoders for neuron classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 197–205, 2024.

[25] Albert Lin, Runzhe Yang, Sven Dorkenwald, et al. Network statistics of the whole-brain connectome of Drosophila. *Nature*, 634(8032):153–165, 2024.

[26] Arie Matsliah, Szi-chieh Yu, Krzysztof Kruk, Doug Bland, et al. Neuronal parts list and wiring diagram for a visual system. *Nature*, 634(8032):166–180, 2024.

[27] Benjamin J. Matthews, Michelle E. Kim, John J. Flanagan, Daisuke Hattori, James C. Clemens, S. Lawrence Zipursky, and Wesley B. Grueber. Dendrite self-avoidance is controlled by Dscam. *Cell*, 129(3):593–604, 2007. doi: 10.1016/j.cell.2007.04.013.

[28] Geoffrey W. Meissner, Aljoscha Nern, Zachary Dorman, Gina M. DePasquale, Kaitlyn Forster, et al. A searchable image resource of Drosophila GAL4 driver expression patterns with single neuron resolution. *eLife*, 12:e80660, 2023. doi: 10.7554/eLife.80660.

[29] MICrONS Consortium. Functional connectomics spanning multiple areas of mouse visual cortex. *Nature*, 640(8058):435–447, 2025.

[30] Alan Peters and M. L. Feldman. The projection of the lateral geniculate nucleus to area 17 of the rat cerebral cortex. I. General description. *Journal of Neurocytology*, 5(1):63–84, 1976.

[31] Ben Piazza, Dániel L. Barabási, Giulia Menichetti, André Ferreira Castro, and Albert-László Barabási. Physical Network Constraints Define the Multiplicative Architecture of the Brain's Connectome. bioRxiv, 2025. URL `https://doi.org/10.1101/2025.02.27.640551`. bioRxiv preprint.

[32] Stephen M. Plaza, Jody Clements, Tom Dolafi, Lowell Umayam, et al. neuPrint: An open access tool for EM connectomics. *Frontiers in Neuroinformatics*, 16:896292, 2022. doi: 10.3389/fninf.2022.896292.

[33] Alexander Ratzan, Sidharth Goel, Junhao Wen, Christos Davatzikos, and Erdem Varol. Predicting functional brain connectivity with context-aware deep neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2933.

[34] Christopher L. Rees, Keivan Moradi, and Giorgio A. Ascoli. Weighing the evidence in Peters' rule: does neuronal morphology predict connectivity? *Trends in Neurosciences*, 40(2):63–71, 2017.

[35] Michael W. Reimann, James G. King, Eilif B. Muller, Srikanth Ramaswamy, and Henry Markram. An algorithm to predict the connectome of neural microcircuits. *Frontiers in Computational Neuroscience*, 9:120, 2015.

[36] Oren Richter and Elad Schneidman. Building the connectome of a small brain with a simple stochastic developmental generative model. *Proceedings of the National Academy of Sciences*, 122(47):e2504913122, 2025. doi: 10.1073/pnas.2504913122.

[37] Alexander Rives, Joshua Meier, Tom Sercu, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[38] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

[39] Emily C. Sales, Emily L. Heckman, Timothy L. Warren, and Chris Q. Doe. Regulation of subcellular dendritic synapse specificity by axon guidance cues. *eLife*, 8:e43478, 2019. doi: 10.7554/eLife.43478.

[40] Joshua R. Sanes and S. Lawrence Zipursky. Synaptic specificity, recognition molecules, and assembly of neural circuits. *Cell*, 181(3):536–556, 2020.

[41] Louis K. Scheffer, C. Shan Xu, Michal Januszewski, et al. A connectome and analysis of the adult Drosophila central brain. *eLife*, 9:e57443, 2020.

[42] Philipp Schlegel, Yijie Yin, Alexander S. Bates, Sven Dorkenwald, et al. Whole-brain annotation and multi-connectome cell typing of Drosophila. *Nature*, 634(8032):139–152, 2024.

[43] Gregory Schwartzman, Ben Jourdan, David García-Soriano, and Arie Matsliah. NTAC: neuronal type assignment from connectivity. *Nature Communications*, 17:1284, 2026. doi: 10.1038/s41467-025-68044-1.

[44] Ruggero Scorcioni, Sridevi Polavaram, and Giorgio A. Ascoli. L-Measure: a web-accessible tool for the analysis, comparison and search of digital reconstructions of neuronal morphologies. *Nature Protocols*, 3(5):866–876, 2008.

[45] Alexander Shapson-Coe, Michał Januszewski, Daniel R. Berger, et al. A petavoxel fragment of human cerebral cortex reconstructed at nanoscale resolution. *Science*, 384(6696):eadk4858, 2024.

[46] Gordon M. G. Shepherd, Armen Stepanyants, Ingrid Bureau, Dmitri Chklovskii, and Karel Svoboda. Geometric and functional organization of cortical circuits. *Nature Neuroscience*, 8 (6):782–790, 2005.

[47] D. A. Sholl. Dendritic organization in the neurons of the visual and motor cortices of the cat. *Journal of Anatomy*, 87(Pt 4):387–406, 1953.

[48] Armen Stepanyants and Dmitri B. Chklovskii. Neurogeometry and potential synaptic connectivity. *Trends in Neurosciences*, 28(7):387–394, 2005.

[49] Armen Stepanyants, Patrick R. Hof, and Dmitri B. Chklovskii. Geometry and structural plasticity of synaptic connectivity. *Neuron*, 34(2):275–288, 2002.

[50] Fabian Svara, Dominique Förster, Fumi Kubo, Michał Januszewski, Marco dal Maschio, Philipp J. Schubert, Jörgen Kornfeld, Adrian A. Wanner, Eva Laurell, Winfried Denk, and Herwig Baier. Automated synapse-level reconstruction of neural circuits in the larval zebrafish brain. *Nature Methods*, 19(11):1357–1366, 2022. doi: 10.1038/s41592-022-01621-0. URL `https://doi.org/10.1038/s41592-022-01621-0`.

[51] Seth R Taylor, Gabriel Santpere, Alexis Weinreb, Alec Barrett, Molly B Reilly, Chuan Xu, Erdem Varol, Panos Oikonomou, Lori Glenwinkel, Rebecca McWhirter, et al. Molecular topography of an entire nervous system. *Cell*, 184(16):4329–4347, 2021.

[52] Jakob Troidl, Johannes Knittel, Wanhua Li, Fangneng Zhan, Hanspeter Pfister, and Srinivas Turaga. Global neuron shape reasoning with point affinity transformers. *bioRxiv*, 2024. 2024.11.24.625067.

[53] Daniel Udvary, Philipp Harth, Jakob H. Macke, Hans-Christian Hege, Christiaan P. J. de Kock, Bert Sakmann, and Marcel Oberlaender. The impact of neuron morphology on cortical network architecture. *Cell Reports*, 39(2):110677, 2022.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[55] Guojia Wan, Minghui Liao, Dong Zhao, Zengmao Wang, Shirui Pan, and Bo Du. Inferring neuron-level brain circuit connection via graph neural network amidst small established connections. *bioRxiv*, 2023. 2023.06.29.547138.

[56] Marissa A. Weis, Laura Hansel, Timo Lüddecke, and Alexander S. Ecker. Self-supervised graph representation learning for neuronal morphologies. *Transactions on Machine Learning Research*, 2023. arXiv:2112.12482.

[57] John Graham White, Eileen Southgate, J. N. Thomson, and Sydney Brenner. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 314(1165):1–340, 1986. doi: 10.1098/rstb.1986.0056. URL `https://doi.org/10.1098/rstb.1986.0056`.

[58] Michael Winding, Benjamin D. Pedigo, Christopher L. Barnes, et al. The connectome of an insect brain. *Science*, 379(6636):eadd9330, 2023.

[59] Daniel Witvliet, Ben Mulcahy, James K. Mitchell, Yaron Meirovitch, Daniel R. Berger, Yuelong Wu, Yufang Liu, Wan Xian Koh, Rajeev Parvathala, Douglas Holmyard, Richard L. Schalek, Nir Shavit, Andrew D. Chisholm, Jeff W. Lichtman, Aravinthan D. T. Samuel, and Mei Zhen. Connectomes across development reveal principles of brain maturation. *Nature*, 596(7871):257–261, 2021.

[60] Szi-Chieh Yu, J. Alexander Bae, Arie Matsliah, Sven Dorkenwald, et al. New synapse detection in the whole-brain connectome of Drosophila. *bioRxiv*, 2025. doi: 10.1101/2025.07.11.664377.

[61] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021.

[62] Zhihao Zheng, J. Scott Lauritzen, Eric Perlman, Camenzind G. Robinson, et al. A complete electron microscopy volume of the brain of adult Drosophila melanogaster. *Cell*, 174(3):730–743.e22, 2018. doi: 10.1016/j.cell.2018.06.019.

# A Data provenance and preprocessing

**FlyWire release.** We use the FlyWire `v783` (Princeton) adult *Drosophila* release [10, 42], reconstructed from the FAFB electron-microscopy volume [62], with the Princeton synapse and connection tables [60]. We download skeletons, soma coordinates, anatomical superclasses, directed connections, and synapse locations once from the FlyWire Codex and Princeton data releases and cache them locally. The intersection of these files contains $138{,}584$ neurons and $3{,}732{,}460$ directed connections after the standard threshold of at least five synapses. We perform no raw-EM processing.

Table 4: FlyWire files and their use. Identity annotations never enter the learned morphology models; they are used only for splits, diagnostic strata, and rule-based controls.

| File | Source | Use |
|---|---|---|
| `sk_lod1_783_healed/` | FlyWire Codex | SWC tokens and morphometrics |
| `coordinates.csv` | FlyWire Codex | soma-neighborhood sampling |
| `classification.csv` | FlyWire Codex | Region split |
| `connections_princeton.csv` | Princeton release | directed labels ($\geq 5$ synapses) |
| `fafb_v783_princeton_synapse_table.csv` | Princeton release | token-localization labels |

**Deterministic preprocessing.** We root each SWC at its soma and decompose it into maximal non-branching segments. To associate synapses with these segments, we group synapse rows by directed neuron pair and project each location to the nearest skeleton vertex on the presynaptic and postsynaptic sides. The connection threshold, split manifest, token cache, and evaluation-pair manifests are fixed before model comparison. Neurons above $2{,}048$ branch tokens are excluded rather than truncated: $5{,}674/138{,}584 = 4.1\%$ are excluded, leaving $132{,}910$ neurons whose complete natural sequences are available to the within-neuron encoder.

# B Tokenization and model details

## B.1 Branch-segment tokens

Breadth-first traversal begins at the soma after the SWC has been converted to a rooted segment tree. BFS supplies a deterministic serialization; it is not a causal modeling assumption. Each segment is represented by the 18-D vector in Table 5. Coordinates are centered separately for each neuron at its soma and normalized by that neuron's extent, so they describe where a segment lies within its own arbor rather than where the soma lies in the brain.

Table 5: Branch-segment token channels. The linear projection consumes all 18 channels jointly. BFS depth and sibling rank are also supplied as learned categorical embeddings after projection.

| Family | Channels and interpretation |
|---|---|
| Orientation (3) | Parent-relative unit direction in a local Gram–Schmidt frame. |
| Geometry and tree context (12) | Segment length, tortuosity, Strahler order, branch angle, child-angle ratio, torsion, sibling count, geodesic depth, bifurcation indicator, subtree size, parent-edge length, and radius. |
| Midpoint (3) | Soma-centered, extent-normalized Cartesian midpoint $(x, y, z)$. |

The 18-D vector is projected to $d = 128$. We then add two learned lookup vectors: one indexed by BFS depth (clipped at 32; a $33 \times 128$ table) and one indexed by sibling rank at the parent bifurcation (clipped at 8; a $9 \times 128$ table). These *tree-position embeddings* are optimized with the rest of the model. They encode rooted-tree context; they are distinct from the three continuous midpoint channels. Variable-length batches are padded to the longest retained sequence in the batch, and padding is removed from every self- and cross-attention operation by a binary mask.

The multiscale token recipe adds up to 16 deterministic subtree summaries. Each summary is a cable-length-weighted mean of descendant segment features at a selected tree anchor. These summaries provide broader arbor context; they are distinct from the 25-D whole-neuron conditioning vector. The matched 79-D MLP pools the same multiscale token channels.

### B.2 Global summary features

The optional 25-D whole-neuron summary contains 11 global morphometrics (including cable length, branch and terminal counts, depth, Strahler order, segment-length moments, bifurcation-angle moments, and tortuosity), eight position pools (extent, centroid offset, and compactness), and six Sholl crossing counts [47]. The first 19 dimensions summarize information represented locally in tokens. To compute a Sholl count, we join the endpoints of every edge and count crossings of a soma-centered sphere. This count depends on the whole arbor and is therefore unavailable from an individual segment token. Conditioning is therefore a deliberate source of global information and an optimization convenience, not a claim that the Transformer must reconstruct every global statistic.

For the matched MLP, we concatenate this summary with the mean, maximum, and standard deviation of each of the 18 token channels, yielding $25+54 = 79$ dimensions per neuron. Connectoformer receives the same 25-D summary and the unpooled sequence carrying those 18 channels. Thus both architectures see the same underlying channels, while only Connectoformer retains segment-level arrangement and partner-conditioned interaction.

### B.3 Encoder and scorer tensor path

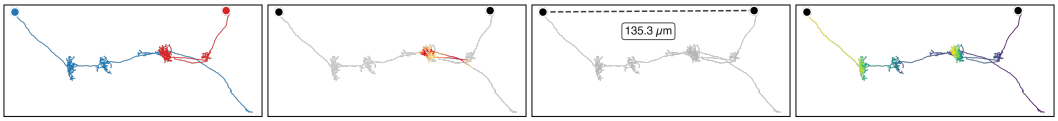



Figure 7: **Four sources of connectivity evidence on a connected pair (Eq. 1), from left to right.** $\mathcal{S}_{\text{global}}$**:** full-color neurons (A blue, B red) illustrate whole-neuron shape compatibility. $\mathcal{S}_{\text{local}}$**:** saturated color in the interaction zone illustrates compatibility between specific branches of the two neurons. $\mathcal{S}_{\text{distance}}$**:** marked somas and their separation illustrate geometric proximity; the arbor prior additionally measures co-occupancy along the arbors. $\mathcal{S}_{\text{tree}}$**:** color indicates tree position through BFS depth and sibling rank. These panels illustrate the conceptual framing; the implemented scorer combines pooled and cross-attention paths.

A shared six-layer pre-norm Transformer (4 heads, $d = 128$, FFN dimension 256, GELU, dropout 0.1) independently maps each neuron to $H_a \in \mathbb{R}^{K_a \times 128}$. Attention is bidirectional because a BFS serialization has no semantic past or future. If summary conditioning is enabled, a small MLP projects the 25-D vector to 128 dimensions and broadcast-adds it to the tokens of that neuron. The two neurons do not interact inside this trunk, so an encoded neuron can be cached and reused across pairs.

For whole-neuron scoring, we use four learned attention-pooled slots over all hidden states rather than the encoder's separate last/mean summary. Source and target slot projections score all 16 slot pairs bilinearly, and log-mean-exp aggregates those scores. These per-neuron slots are partner-independent. The local path approximately uniformly retains at most $K = 256$ encoded tokens per neuron. Queries from $a$ attend to keys/values from $b$; queries from $b$ then attend to the updated source states in the bidirectional cross-attention step. Separate 16-D pre- and postsynaptic projections form a $K_a \times K_b$ compatibility matrix. Log-mean-exp over valid matrix entries gives the directed pair logit. For localization, we aggregate the five strongest partner-token matches in each row or column into a presynaptic or postsynaptic score, respectively. A learned scalar $\alpha = \sigma(\alpha_{\text{logit}})$, initialized at 0.5, blends pooled and local logits. Therefore $\alpha$ is learned, not tuned as a fixed hyperparameter.

Writing the two path logits as $s_{\text{pooled}}$ and $s_{\text{cross}}$, the morphology logit is

$$s_{\text{morph}}(a, b) = \alpha s_{\text{pooled}}(a, b) + (1 - \alpha) s_{\text{cross}}(a, b), \qquad \alpha = \sigma(\alpha_{\text{logit}}). \tag{4}$$

Learned-overlap fusion additionally consumes the overlap feature as described in Appendix D.

**Optimization.** The Transformer uses AdamW at learning rate $3 \times 10^{-4}$, weight decay $10^{-4}$, a 20-epoch linear warmup, cosine decay, bfloat16 activations, and at most 1,000 epochs. Checkpoint selection uses validation metrics only. For Region, the validation set is the visual projection and visual centrifugal neurons, held out from training, rather than a random subset of the training region (Section 3). Dense-subset training encodes a high-edge-density subset of up to 10,240 neurons and samples 16,384 positive and 16,384 negative pairs sharing those endpoints. Reusing each neuron's

encoding across these pairs reduces the encoding cost. We match the MLP's exposure to training pairs, but not its parameter count or floating-point operations.

MLP baselines use Adam with recipe-specific epoch limits and validation-based selection; the matched 79-D control allows up to $1{,}000$ epochs and the summary-only control up to $500$. The headline Transformer objective includes connectivity BCE, soft node-distribution loss with weight $0.02$, and margin-ranking loss with weight $0.1$. Node targets are derived from training synapses, not supplied as inference features. The BCE-only and auxiliary-loss variants in Appendix J isolate their effects.

# C Evaluation protocols and statistics

**Reproducibility.** Code for all experiments is available at `https://github.com/sgoe1/connectoformer`. The appendices specify data provenance, features, training, evaluation panels, and uncertainty units.

## C.1 Metric construction

- **Overall AUROC/AP** uses a balanced evaluation panel whose negatives are half uniformly drawn and half soma-neighborhood drawn. These are discrimination metrics, not population-calibrated precision.
- **Hard AUROC/AP** compares positives with non-connections drawn from the source neuron's soma-nearest neighborhood. It suppresses gross distance as an easy cue.
- **Overlap Hard AUROC/AP** compares positives with non-connections whose voxelized arbors overlap above the median. It tests ranking after conditioning on physical opportunity, and is reported only in Appendix L.1.
- **In-Type AUROC/AP** requires each negative pair to share a neuroblast group and lie inside the soma-neighborhood band. It controls a coarse developmental identity cue more tightly than Hard, but not exact annotated source and target type.
- **Directed AUROC** scores $a \to b$ against $b \to a$ for connected ordered pairs. **Weak-edge AUROC** evaluates pairs with one to four synapses, below the primary five-synapse threshold. Both are reported only in Appendix L.1.

These ranking metrics do not establish probability calibration. Post-hoc temperature scaling [13] is a separate, validation-fitted procedure; we do not claim a temperature-scaling result here.

Random uses five independently generated four-fold neuron-disjoint partitions. Table 2 and the complete results below report SE over five seed-level units, and paired inference uses the same units: Random first averages the four folds within each partition seed. Region has one fixed optic-lobe holdout and five optimization seeds, with checkpoints selected on held-out visual projection and centrifugal neurons; Patch has five seeded partitions. Lineage and shape-cluster transfer use five held-out-group folds and one fixed optimization seed. We compute SD and paired effects across these folds. We use two-sided paired $t$ intervals and Holm correction over the seven architecture contrasts in Table 8, separately by split and endpoint.

Table 6: Evaluation protocols and replicate units. Random, Region, and Patch report mean $\pm$ SE over five seed-level units, which are also the units for paired tests; group transfer reports SD across its five held-out-group folds, which are its paired units.

| Protocol | Held out | Runs | Replicate unit (what varies) |
|---|---|---|---|
| Random | neurons | 5 seeds $\times$ 4 folds | seed, folds averaged (partition, optimization) |
| Region | optic-lobe region | 5 seeds | seed (optimization) |
| Patch | seeded spatial partition | 5 seeds | seed (partition, optimization) |
| Lineage | developmental groups | 5 folds | fold (held-out group) |
| Shape cluster | morphology clusters | 5 folds | fold (held-out group) |

## C.2 Complete results

Table 7 expands Table 2 with results for the learned-overlap condition. All rows share the evaluation panels, replicate units, and validation-based checkpoint selection of Table 2.

Table 7: Complete AUROC results, including the learned-overlap condition. Mean ± SE over five replicate units as in Table 2. These are the same runs as Table 2; AP is not reported for them.

| | Random | | | Region | | | Patch | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **Overall** | **Hard** | **In-Type** | **Overall** | **Hard** | **In-Type** | **Overall** | **Hard** | **In-Type** |
| Hemilineage | .519 ± .001 | .434 ± .002 | .171 ± .002 | .537 ± .000 | .488 ± .000 | .242 ± .000 | .517 ± .014 | .404 ± .019 | .157 ± .018 |
| Soma proximity | .479 ± .000 | .050 ± .001 | .047 ± .001 | .483 ± .000 | .056 ± .000 | .057 ± .000 | .492 ± .007 | .090 ± .008 | .087 ± .008 |
| Arbor overlap | .841 ± .001 | .686 ± .002 | .569 ± .002 | .834 ± .000 | .668 ± .000 | .605 ± .000 | .864 ± .003 | .728 ± .006 | .595 ± .017 |
| MLP (54-D token statistics) | .787 ± .002 | .767 ± .003 | .759 ± .003 | .671 ± .003 | .659 ± .003 | .657 ± .003 | .765 ± .007 | .733 ± .008 | .710 ± .009 |
| MLP (25-D summary) | .839 ± .001 | .788 ± .002 | .770 ± .002 | .692 ± .002 | .689 ± .003 | .681 ± .005 | .812 ± .005 | .748 ± .006 | .713 ± .006 |
| MLP (79-D combined) | .841 ± .002 | .789 ± .003 | .770 ± .003 | .694 ± .010 | .692 ± .009 | .688 ± .013 | .809 ± .005 | .743 ± .005 | .705 ± .005 |
| Connectoformer (tokens only) | .823 ± .001 | .812 ± .002 | .809 ± .002 | .702 ± .003 | .701 ± .004 | .700 ± .004 | .798 ± .005 | .779 ± .007 | .761 ± .006 |
| Connectoformer + summary | .862 ± .001 | .842 ± .001 | .834 ± .002 | .677 ± .010 | .673 ± .013 | .663 ± .015 | .824 ± .003 | .790 ± .006 | .769 ± .004 |
| MLP (25-D summary) + arbor prior | .909 ± .001 | .823 ± .002 | .804 ± .002 | .852 ± .011 | .709 ± .023 | .706 ± .026 | .909 ± .005 | .825 ± .011 | .785 ± .010 |
| MLP (79-D combined) + arbor prior | .911 ± .001 | .825 ± .002 | .806 ± .001 | .850 ± .005 | .706 ± .011 | .704 ± .011 | .911 ± .005 | .827 ± .011 | .787 ± .011 |
| Connectoformer + summary + arbor prior | .916 ± .001 | .837 ± .002 | .826 ± .002 | .867 ± .003 | .745 ± .004 | .751 ± .005 | .913 ± .005 | .833 ± .011 | .802 ± .010 |
| MLP (79-D combined) + learned overlap | .941 ± .000 | .883 ± .001 | .818 ± .001 | .897 ± .001 | .796 ± .002 | .761 ± .003 | .943 ± .001 | .886 ± .004 | .787 ± .007 |
| Connectoformer + summary + learned overlap | .957 ± .000 | .919 ± .000 | .896 ± .001 | .891 ± .003 | .794 ± .004 | .782 ± .004 | .948 ± .002 | .898 ± .006 | .835 ± .008 |

Table 8: Paired architecture effects for the runs in Table 2 and Table 10. Values are mean AUROC difference with two-sided $t$-based 95% CI over five paired units: Random folds are averaged within partition seed; Region and Patch units are seeds. Intervals are unadjusted; the text states which effects survive Holm correction across these seven contrasts.

| **Contrast** | **Split** | **Overall** | **Hard** | **In-Type** |
|---|---|---|---|---|
| Conn. + summary − 79-D MLP | Random | $+.021$ $[+.018, +.025]$ | $+.053$ $[+.048, +.058]$ | $+.065$ $[+.060, +.070]$ |
| | Region | $-.016$ $[-.049, +.016]$ | $-.019$ $[-.056, +.017]$ | $-.025$ $[-.071, +.021]$ |
| | Patch | $+.015$ $[+.002, +.028]$ | $+.047$ $[+.027, +.068]$ | $+.064$ $[+.044, +.083]$ |
| + arbor prior: Conn. − 25-D MLP | Random | $+.006$ $[+.005, +.008]$ | $+.015$ $[+.011, +.019]$ | $+.022$ $[+.017, +.026]$ |
| | Region | $+.015$ $[-.019, +.048]$ | $+.036$ $[-.035, +.107]$ | $+.045$ $[-.032, +.122]$ |
| | Patch | $+.004$ $[+.000, +.007]$ | $+.008$ $[+.003, +.014]$ | $+.017$ $[+.012, +.022]$ |
| + arbor prior: Conn. − 79-D MLP | Random | $+.005$ $[+.004, +.007]$ | $+.013$ $[+.008, +.017]$ | $+.019$ $[+.013, +.025]$ |
| | Region | $+.016$ $[-.002, +.035]$ | $+.039$ $[+.001, +.077]$ | $+.047$ $[+.008, +.086]$ |
| | Patch | $+.003$ $[-.001, +.007]$ | $+.007$ $[+.001, +.012]$ | $+.015$ $[+.010, +.020]$ |
| + learned overlap: Conn. − 79-D MLP | Random | $+.017$ $[+.016, +.017]$ | $+.036$ $[+.034, +.037]$ | $+.079$ $[+.077, +.080]$ |
| | Region | $-.006$ $[-.014, +.002]$ | $-.002$ $[-.016, +.012]$ | $+.020$ $[+.003, +.038]$ |
| | Patch | $+.005$ $[-.001, +.011]$ | $+.012$ $[-.000, +.024]$ | $+.048$ $[+.028, +.068]$ |
| Tokens only − 54-D MLP | Random | $+.036$ $[+.033, +.039]$ | $+.044$ $[+.042, +.047]$ | $+.049$ $[+.046, +.053]$ |
| | Region | $+.031$ $[+.027, +.036]$ | $+.042$ $[+.034, +.050]$ | $+.043$ $[+.035, +.051]$ |
| | Patch | $+.033$ $[+.016, +.049]$ | $+.046$ $[+.033, +.059]$ | $+.051$ $[+.035, +.066]$ |
| Tokens only − 25-D MLP | Random | $-.016$ $[-.020, -.012]$ | $+.024$ $[+.021, +.026]$ | $+.039$ $[+.037, +.040]$ |
| | Region | $+.010$ $[+.005, +.015]$ | $+.012$ $[+.003, +.022]$ | $+.019$ $[+.005, +.032]$ |
| | Patch | $-.014$ $[-.035, +.007]$ | $+.030$ $[+.007, +.054]$ | $+.048$ $[+.027, +.070]$ |
| Tokens only − 79-D MLP | Random | $-.018$ $[-.023, -.013]$ | $+.023$ $[+.018, +.027]$ | $+.039$ $[+.034, +.044]$ |
| | Region | $+.008$ $[-.022, +.038]$ | $+.009$ $[-.020, +.039]$ | $+.012$ $[-.026, +.050]$ |
| | Patch | $-.011$ $[-.030, +.008]$ | $+.036$ $[+.015, +.057]$ | $+.056$ $[+.037, +.074]$ |

For the paired effects in Table 8, Holm correction is applied across the seven contrasts separately for each split and metric, using $p$-values from the unrounded seed-level differences. On Random, every effect survives, including the summary MLPs' overall advantage over tokens-only Connectoformer. On Patch, every In-Type gain survives, as do the Hard gains of summary-conditioned and tokens-only Connectoformer over the 79-D MLP and all tokens-only versus 54-D MLP effects. On Region, only the tokens-only versus 54-D MLP effects and the tokens-only versus 25-D MLP overall effect survive; no summary-conditioned Connectoformer contrast does, although the arbor-prior contrast against the 79-D MLP favors Connectoformer in all five seeds.

### C.3 Group-disjoint paired effects

Table 9 complements the absolute fold SDs in Table 3 with paired AUROC and AP effects. The five held-out groups are the inferential units. Shape-cluster holdout is the clearest architecture result, particularly when overlap is learned jointly; lineage holdout supports gains on the same-neuroblast In-Type subset, while the Hard effect remains imprecise.

Table 9: Five-fold group transfer, Connectoformer minus matched MLP. Each cell is mean difference [95% CI].

| Input | Held out | Metric | Overall | Hard | In-Type |
|---|---|---|---|---|---|
| Morphology | Lineage | AUROC | $-.0093[-.0522, .0335]$ | $+.0464[-.0260, .1187]$ | $+.0347[-.0221, .0915]$ |
| | | AP | $-.0017[-.0385, .0351]$ | $+.0134[-.0176, .0444]$ | $+.0141[-.0497, .0779]$ |
| | Shape | AUROC | $-.0002[-.0239, .0235]$ | $+.0389[-.0004, .0781]$ | $+.0384[.0150, .0618]$ |
| | | AP | $+.0034[-.0287, .0356]$ | $+.0416[-.0161, .0993]$ | $+.0451[.0108, .0794]$ |
| + learned overlap | Lineage | AUROC | $-.0065[-.0108, -.0022]$ | $+.0219[-.0013, .0451]$ | $+.0463[.0090, .0836]$ |
| | | AP | $-.0030[-.0050, -.0010]$ | $+.0090[-.0015, .0194]$ | $+.0612[.0252, .0971]$ |
| | Shape | AUROC | $+.0095[.0038, .0152]$ | $+.0307[.0088, .0525]$ | $+.0638[.0187, .1089]$ |
| | | AP | $+.0166[.0073, .0259]$ | $+.0360[.0112, .0608]$ | $+.0906[.0329, .1482]$ |

## D Overlap and component controls

### D.1 Learned overlap versus a fixed feasibility prior

At $128^3$ resolution, the overlap scalar $o(a, b) \in [0, 1]$ is computed once from voxelized arbor co-occupancy and cached. Overlap alone reaches Random Hard/In-Type AUROC $0.686/0.569$, whereas the morphology summary MLP reaches $0.788/0.770$. Physical opportunity is therefore useful but not sufficient for partner ranking.

The primary control gives every learned scorer a detached feature $\log(1 + o/10^{-4})$. A two-layer, width-16 residual MLP jointly consumes that feature and the architecture's morphology logit(s); its final layer is zero-initialized, so training starts exactly at the morphology-only score. This is the *learned-overlap* condition. It allows the optimizer to downweight, invert, or interact with overlap rather than imposing one-sided behavior.

Table 10 summarizes the matched-input comparison on all three splits. Learning overlap improves overall AUROC for both scorers on every split. Connectoformer's advantage over the matched MLP persists on every Random metric and on Patch In-Type. On Region it ties on overall and Hard AUROC and leads on In-Type only before multiplicity correction ($+0.020$ $[+0.003, +0.038]$); on Patch its Hard gain ($+0.012$ $[+0.000, +0.024]$) is not significant after Holm correction. Complete AUROC results are in Table 7; Table 8 reports the paired tests.

Table 10: **Learning overlap jointly.** Both scorers receive morphology and the same overlap scalar; neither uses the fixed penalty. AUROC mean with SE subscript over five replicate units, as in Table 2. **Bold**: numerical maxima, not paired significance.

| | Random | | | Region | | | Patch | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **Overall** | **Hard** | **In-Type** | **Overall** | **Hard** | **In-Type** | **Overall** | **Hard** | **In-Type** |
| MLP (79-D combined) + learned overlap | $.941_{\pm.000}$ | $.883_{\pm.001}$ | $.818_{\pm.001}$ | $\mathbf{.897_{\pm.001}}$ | $\mathbf{.796_{\pm.002}}$ | $.761_{\pm.003}$ | $.943_{\pm.001}$ | $.886_{\pm.004}$ | $.787_{\pm.007}$ |
| Connectoformer + summary + learned overlap | $\mathbf{.957_{\pm.000}}$ | $\mathbf{.919_{\pm.000}}$ | $\mathbf{.896_{\pm.001}}$ | $.891_{\pm.003}$ | $.794_{\pm.004}$ | $\mathbf{.782_{\pm.004}}$ | $\mathbf{.948_{\pm.002}}$ | $\mathbf{.898_{\pm.006}}$ | $\mathbf{.835_{\pm.008}}$ |

The fixed arbor prior instead applies

$$s'(a, b) = s(a, b) + \beta \min\left(0, \log \frac{\max(o(a, b), \epsilon)}{\tau}\right), \qquad \epsilon = 10^{-8}. \tag{5}$$

At the default setting $\beta = 1, \tau = 10^{-4}$, the score is unchanged when overlap exceeds $\tau$ and is reduced when overlap is very low. This is an explicit feasibility filter, not learned evidence credited to either architecture. To test sensitivity to $\beta$ and $\tau$, we apply the prior post hoc to saved morphology logits from earlier checkpoints of both architectures, not the runs in Table 2, on six paired Random partitions spanning five partition seeds. Over a $5\times5$ grid, Hard AUROC stays within $0.881$–$0.885$ for positive $\beta$, and Connectoformer exceeds the matched MLP in all 25 cells. At the default setting the

seed-level Hard AUROC/AP effects are $+.0409$ $[+.0313, +.0505]$ and $+.0298$ $[+.0243, +.0353]$. Uniform-negative AUROC is nearly saturated, so the repeated learned-overlap training grid, not this fixed-checkpoint sensitivity analysis, supports the architecture comparison.

### D.2 Partner retrieval examples

The panels above score fixed sets of positive and negative pairs. To show what the scores look like when a single neuron's partners must be found among all candidates, we use the Connectoformer + summary + learned-overlap checkpoint from Random partition seed $42$, fold $0$, and score held-out presynaptic neurons with at least three held-out postsynaptic partners against all $33{,}233$ held-out neurons. No training or validation neuron is a candidate. This is inference on a fixed checkpoint, without retraining.

Figure 8 shows four of these sources, chosen to illustrate both directions of disagreement with raw arbor overlap. In (a) and (b) the model ranks a true partner first where overlap ranks the first partner 105th and 54th. In (c) and (d) overlap ranks a partner first, while the model places large wide-field neurons (Dm, Pm and Li types) above every partner.

### D.3 Retrained and fixed-checkpoint ablations

To measure the contribution of the global summary, we compare models retrained with and without it. In a separate Random panel with the same 20 held-out-fold structure, a tokens-only Connectoformer reaches $0.831 \pm .009/0.820 \pm .009/0.821 \pm .006$ Overall/Hard/In-Type. Its within-panel summary-conditioned counterpart reaches $0.876 \pm .005/0.853 \pm .008/0.850 \pm .005$. In Table 2, the 25-D summary MLP reaches $0.839/0.788/0.770$ and tokens-only Connectoformer $0.823/0.812/0.809$. Thus the summary primarily improves broad ranking, while tokens alone already exceed the summary MLP on the two hard strata. Pooling token channels does not recover the hard-case performance of the branch-token model.

Table 11 collects the retraining ablations. Removing dense-subset training leaves overall AUROC unchanged but lowers Hard AUROC by about five points; removing tree-position embeddings costs less than one point. A causal BFS attention mask changes almost nothing, consistent with BFS order having no causal direction, and training without hard-negative sampling raises overall AUROC to .903 (its Hard AUROC was not recorded). These panels come from earlier model generations than Table 2, so only differences within a block are interpretable.

Table 11: **Which components improve connectivity prediction?** Each block is a separate retraining panel with its own reference row and Random protocol; compare variants within a block, not across blocks or with Table 2. Both panels predate the runs in Table 2 and report each run's last logged evaluation rather than a validation-selected checkpoint. First block: mean ± SD over 20 fold runs; second: mean ± SE over five seeds of an 80/20 neuron holdout without a validation split. The last two rows report means only; Hard AUROC was not recorded for the run without hard-negative sampling.

| Variant | Overall AUROC | Hard AUROC |
|---|---|---|
| *Global summary: separate 20-fold retraining panel* | | |
| Connectoformer + summary (reference) | $.876 \pm .005$ | $.853 \pm .008$ |
| Without global summary | $.831 \pm .009$ | $.820 \pm .009$ |
| *Training and tree context: five-seed ablation panel* | | |
| Bidirectional trunk (reference) | $.882 \pm .003$ | $.845 \pm .002$ |
| Without tree-position embeddings | $.878 \pm .002$ | $.838 \pm .004$ |
| Without dense-subset training | $.882 \pm .003$ | $.799 \pm .002$ |
| Causal BFS attention mask | .882 | .844 |
| Without hard-negative sampling | .903 | — |

We separately zero-mask each input family at inference to test the trained checkpoint's dependence on those features. Because masking changes the input distribution, it does not measure what a model retrained without that family could learn. Table 12 reports paired bootstrap intervals on one fixed panel.

The large final row shows that the selected checkpoint relies heavily on its conditioning route. It does not imply that summary information is worth $0.277$ after retraining; the direct retraining gap

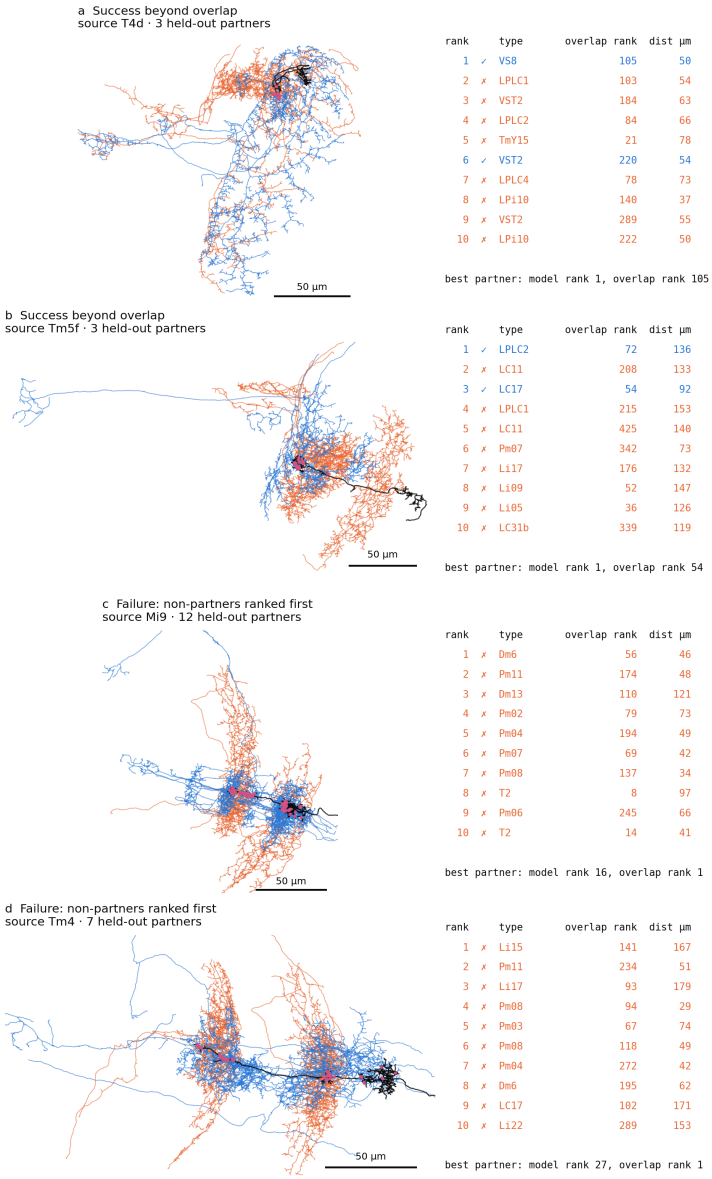



Figure 8: **Ranked partners for four held-out presynaptic neurons.** Connectoformer + summary + learned overlap (Random seed 42, fold 0) ranks all 33,233 held-out neurons as candidate partners of each source. **Left:** frontal projection of the source (black), its held-out postsynaptic partners (blue), up to four non-partners from the model's top ten (orange), and the source's presynaptic sites onto the shown partners (magenta). **Right:** the model's top ten candidates, marked as connected (blue, check) or not connected (orange, cross), with each candidate's rank under raw arbor overlap, soma–soma distance ($\mu$m), and FlyWire type. Extended version of Figure 4.

Table 12: Fixed-checkpoint feature-family sensitivity. Negative values mean AUROC decreases after zero-masking.

| **Masked family** | **$\Delta$ AUROC [95% bootstrap CI]** |
|---|---|
| Parent-relative direction | $+.0002$ $[-.0007, +.0012]$ |
| Scalar morphometrics | $-.0158$ $[-.020, -.012]$ |
| Soma-relative midpoint | $-.0041$ $[-.0066, -.0018]$ |
| Tree-position embedding | $-.0654$ $[-.077, -.054]$ |
| 25-D global summary | $-.2772$ $[-.296, -.258]$ |

is much smaller. Six of the summary dimensions are global Sholl counts unavailable to one local token, so some reliance also reflects deliberate non-local information supplied to both architecture families.

### D.4 Shortcut and coordinate diagnostics

An identity-only MLP reaches Overall/Hard/In-Type AUROC $0.788/0.758/0.753$, below the morphology summary MLP's $0.864/0.806/0.798$ in that control panel. Adding identity to morphology changes In-Type by $+.001$. On a stricter immutable panel of $32{,}768$ examples in 128 balanced annotated source–target type pairs, Connectoformer reaches macro AUROC $0.5603$ (95% CI $[0.552, 0.568]$) and the matched MLP $0.5677$ ($[0.557, 0.578]$); the paired difference is $-.0074$ $[-.0159, +.0010]$ ($p = .083$). Morphology retains modest partner-level signal with exact types fixed, but this panel does not support an architectural gain. The panels are not nested: this diagnostic fixes exact annotated types, whereas In-Type jointly matches neuroblast group and soma neighborhood. The reported intervals are descriptive $t$ intervals across type-pair cells; cells can share neurons, so they are not independent training replicates.

Centering coordinates at each soma and normalizing by neuronal extent makes the midpoint channels translation-invariant and extent-normalized. These operations do not make the full model scale- or rotation-invariant. Other channels retain physical size information. On a fixed paired panel, a $180^\circ$ dorsoventral yaw changes AUROC from $0.8749$ to $0.8721$, a difference of $-.0028$ $[-.0046, -.0010]$. The effect is small but measurable, so we describe anatomical-frame sensitivity rather than claim rotation invariance.

### D.5 Learned neuron representations

Although the model receives neither cell-type labels nor brain coordinates, morphology may contain information about both. To examine this possibility, we take the pooled neuron embedding of the morphology-only Connectoformer + summary checkpoint from Random partition seed $42$, fold $0$, and analyze only the $33{,}233$ held-out neurons (Figure 9). As a reference we use the $25$-D morphology summary, standardized on training neurons.

Held-out neurons group by cell identity. Among each neuron's ten nearest neighbors (cosine distance), $41\%$ share its FlyWire primary type, against $34\%$ for the summary and $1\%$ by chance. Hemilineage agreement is $31\%$ versus $27\%$ (chance $1\%$), and super-class agreement is $90\%$ versus $85\%$ (chance $40\%$). Agreement is highest for the many repeated optic-lobe columnar types. We next ask whether nearby embeddings also share connectivity and location. Nearest neighbors share postsynaptic partner types (Jaccard $0.32$ versus $0.28$ for the summary and $0.04$ for a random neuron of the same super-class), but when we restrict the comparison to the nearest neighbor of a *different* type, the overlap falls to $0.08$ (summary $0.07$, random $0.02$). A linear readout of soma position, fit on training neurons, reaches $R^2 = 0.38$ ($0.22$, $0.44$ and $0.48$ along x, y and z; summary $0.24$), yet predicts the hemisphere for only $66\%$ of neurons. These results are consistent with region-level information associated with cell identity, while precise soma localization remains limited.

### D.6 Spatial distribution of held-out connectivity performance

Figure 6 uses frozen summary-conditioned Connectoformer and matched 79-D MLP checkpoints, without an arbor prior or learned overlap. We expand six existing Random test panels: seed/fold $1024/2$, $137/0$, $137/1$, $256/3$, $42/0$, and $512/2$. Each panel includes all eligible held-out positive edges and all held-out nonedges in the source neuron's 64-nearest-soma candidate set, rather than

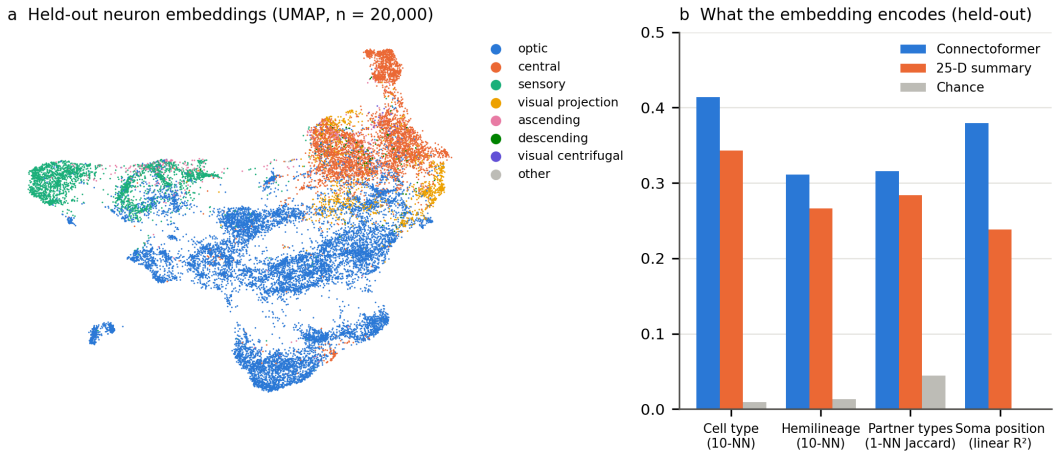



Figure 9: **Held-out neuron embeddings organize by cell identity.** Morphology-only Connectoformer + summary, Random seed 42, fold 0; held-out neurons only. **(a)** UMAP (cosine, 30 neighbors) of 20,000 randomly sampled held-out neurons, colored by FlyWire super-class. **(b)** What the embedding encodes, compared with the 25-D morphology summary and chance: agreement of cell type and hemilineage among each neuron's ten nearest neighbors; Jaccard overlap of postsynaptic partner types with the nearest neighbor (chance: a random neuron of the same super-class); and linear $R^2$ of soma position, which the model never receives. The controls suggest that partner-type sharing and position are partly associated with cell identity (Section D.5).

all possible connectome pairs. Both endpoints are in that checkpoint's test split, excluding training and validation neurons. The panels contain 1,399,553 positive and 3,265,725 hard-negative pair instances; pairs are unique within each panel but can recur across panels. No model is retrained or selected for this analysis. These checkpoints come from an earlier, validation-selected training run of the same two architectures, not the runs in Table 2. Rescoring reproduces the saved scores within numerical tolerance.

We compute AUROC separately in each $25 \times 25\,\mu\mathrm{m}$ XY bin, grouping pairs by source soma for outgoing and target soma for incoming performance, and pooling all z depths. The bin width and minimum support were chosen from anatomy and sample coverage before inspecting spatial model differences. A bin receives a performance color only if every panel contains at least 30 unique positives and 30 unique hard negatives in it; 296 of 363 occupied bins pass for outgoing and 284 for incoming pairs. AUROCs and paired differences are averaged over folds within seed and then equally over the five seeds. The resulting whole-panel Hard AUROCs are $0.860/0.814$; they differ from Table 2 because the checkpoints and panels differ. Spatial bins share neurons and pairs, so the maps are descriptive, without per-bin significance claims or interpolation of unsupported areas. Patterns may reflect neuron-type and partner composition; they do not measure synapse localization or isolate causal regional effects. Saved pair scores, support counts, and provenance allow the visualization to be reproduced without repeating inference.

## E Synapse localization diagnostics

The localization experiments evaluate two related but non-equivalent questions:

- **Pair-conditioned localization AUROC** ranks labeled synapse-bearing tokens against non-synaptic candidates within the sampled pair and candidate set used by the auxiliary loss.
- **Strict Hit@$k$** asks whether any true synaptic token appears among the top $k$ scores when all tokens on that side of the pair are candidates. This is the primary retrieval estimate.

**Full-sequence retrieval.** We score one fixed checkpoint on sampled connected pairs, using synapse labels whose tokenization and neuron-ID order match the token cache. No model was retrained or selected for this analysis. Raw cross-attention scores are reduced by log-sum-exp to one score per token on each side, over the full sequences. The 1,000 sampled rows comprise 994 distinct directed pairs and 1,753 neurons. All rows have labels and finite outputs on both sides.

The pair sampler did not impose the training-time 2,048-token limit: 466 rows contain an above-cap neuron, and sequences reach 32,768 tokens. We retain that full population and report the 534 training-eligible pairs separately, using a subset definition fixed before inspecting the scores. Hit@$k$ is a pair-side success rate, not precision among the $k$ tokens or the fraction of all synapses recovered.

Table 13: Full-sequence localization, in percent. Observed brackets are unadjusted 95% shared-neuron dyadic sensitivity intervals, conditional on the fixed model and panel population. Random columns give exact conditional expectations, not another trained model. Both populations and sides are shown.

| | Presynaptic | | Postsynaptic | |
|---|---|---|---|---|
| **Metric** | **Observed [95%]** | **Random** | **Observed [95%]** | **Random** |
| *Full sampled panel:* 1,000 *pairs* | | | | |
| Hit@1 | 10.0 [8.1, 11.9] | 1.6 | 2.4 [1.5, 3.3] | 1.3 |
| Hit@3 | 21.4 [18.7, 24.1] | 4.7 | 5.7 [4.3, 7.1] | 3.7 |
| Hit@5 | 28.7 [25.7, 31.7] | 7.5 | 8.0 [6.3, 9.7] | 6.0 |
| Hit@10 | 41.0 [37.7, 44.3] | 13.6 | 14.4 [12.2, 16.6] | 11.1 |
| *Training-eligible subset:* 534 *pairs, both neurons* $\leq$ 2,048 *tokens* | | | | |
| Hit@1 | 12.7 [9.8, 15.7] | 2.1 | 2.8 [1.4, 4.2] | 1.6 |
| Hit@3 | 24.9 [21.1, 28.7] | 5.9 | 7.1 [5.0, 9.3] | 4.6 |
| Hit@5 | 32.6 [28.5, 36.6] | 9.3 | 9.4 [6.9, 11.8] | 7.4 |
| Hit@10 | 47.8 [43.4, 52.1] | 16.8 | 17.6 [14.4, 20.8] | 13.8 |

For a side with $N$ candidate tokens and $m$ synaptic tokens, random Hit@$k$ is $1 - \binom{N-m}{k}/\binom{N}{k}$, averaged over the same evaluated rows. This matches both candidate length and positive-label density. Intervals account for shared endpoints through a dyadic variance calculation, grouping duplicate and reciprocal dyads. A 10,000-draw shared-node weighting stress analysis retains the presynaptic advantage: the Hit@10 pre-minus-post contrast is 26.6 percentage points, with dyadic interval $[22.8, 30.4]$ and stress interval $[20.2, 33.0]$. These are dependence sensitivities, not independent training replicates or a causal explanation of the asymmetry.

These rates show presynaptic enrichment over random ranking, with a smaller postsynaptic excess. They do not identify every synapse or demonstrate superiority over another trained localization model.

In the loss ablation, adding synapse-location supervision raises pair-conditioned localization AUROC from 0.493 to 0.711. A separate post-hoc audit restricted to the $K = 256$ scorer-selected candidates gives presynaptic Hit@10 34.4% versus 13.8% under pair-conditional random ordering. Because this audit uses a restricted candidate set and a different protocol, we cannot compare its result directly with the full-sequence retrieval in Table 13. Increasing $K$ from 256 to 768 raises candidate coverage and presynaptic Hit@10 to 40.8%, but changes common-support AUROC by $-.0177$ pre and $-.0833$ post and does not improve postsynaptic Hit@10.

Projection artifacts do not account for the observed presynaptic advantage on the audited panel. Across 1,000 fixed pairs, pre/post AUROC is $0.737/0.632$, for a post-minus-pre difference of $-.106$ $[-.131, -.080]$. The deficit remains $-.077$ to $-.112$ after matching nearest projection error and local segment size. When projection-error p90 is at most 1,000 nm, presynaptic AUROC/Hit@10 remains $0.765/36.4\%$. These controls show that measured mapping artifacts do not fully explain the asymmetry; they do not establish its biological cause.

These localization scores can prioritize branches for inspection, an output our pooled-feature baselines do not provide. They do not establish precise synapse placement. Supervised partner-conditioned placement with compartmental or molecular inputs is a natural next task.

# F Cross-connectome transfer

## F.1 Male Drosophila CNS

We use the `male-cns:v0.9` release, accessed through neuPrint [32], of the male adult *Drosophila* CNS connectome [2]. The FlyWire-trained models are evaluated without optimization on an immutable male-CNS manifest containing 7,139 positives, 4,056 hard negatives, and 4,048 easy neg-

atives. The matched MLP and Connectoformer use the identical pair set. Table 14 separates the cross-individual hard-case result from easy geometry-driven ranking.

Table 14: Zero-shot transfer to the male adult *Drosophila* CNS. The paired confidence intervals compare model scores on the same manifest, not retraining variability.

| Method | Overall AUROC | Hard AUROC | Easy AUROC |
|---|---|---|---|
| Soma proximity | .430 | .018 | .843 |
| Arbor overlap | .809 | .623 | .997 |
| Matched 79-D MLP | .748 | .701 | .795 |
| Connectoformer | .761 | .739 | .783 |
| Connectoformer − MLP | $+.013$ [.008, .018] | $+.038$ [.032, .044] | $-.012$ [$-.017, -.006$] |

The hard-negative gain supports transfer of morphology-based partner discrimination between two flies. The MLP's advantage on easy negatives also shows why this gain should be evaluated separately from geometry-driven ranking.

### F.2 Repeated C. elegans probes

We derive 216 adult worm skeletons and 2,155 directed connections from Dataset 8 of Witvliet et al. [59]. Skeletons and vertex-linked synapses come from the same CATMAID export; polyadic active zones are expanded to directed dyads, and neurons without curated skeletons are removed. The probe uses five independently seeded, five-fold neuron-disjoint outer panels. Hyperparameter selection occurs only inside each outer-training neuron set. Table 15 reports sample SD across the five panel seeds.

Table 15: *C. elegans* transfer probes using a frozen FlyWire-trained Connectoformer checkpoint. Values are mean ± SD across five independently seeded neuron-disjoint CV panels. "10% labels" means that the probe is fit using a random 10% of the labeled pairs available among each outer fold's training neurons; evaluation remains on held-out neurons.

| Probe input | 10% labels AUROC | All labels AUROC |
|---|---|---|
| Raw morphology | $.624 \pm .008$ | $.676 \pm .013$ |
| Frozen FlyWire-pretrained encoding | $.603 \pm .020$ | $.640 \pm .010$ |
| Randomly initialized encoding | $.586 \pm .016$ | $.630 \pm .016$ |
| Soft geometry | $\mathbf{.900 \pm .003}$ | $.914 \pm .005$ |
| Raw morphology + geometry | $.894 \pm .009$ | $\mathbf{.917 \pm .006}$ |
| Pretrained encoding + geometry | $.866 \pm .017$ | $.897 \pm .010$ |

At the locked 10% endpoint, pretrained minus raw is $-.0206$ [$-.0489, +.0077$] and pretrained minus random initialization is $+.0174$ [$-.0078, +.0426$]; neither is positive after Holm correction. Geometry alone reaches 0.900, and pretrained features do not improve it. Separately, the zero-shot FlyWire pair scorer reaches Hard AUROC 0.586, below overlap (0.654), while cosine on pooled encodings is anti-predictive (0.443). Morphology and geometry predict worm connectivity in these tests, but we find no consistent improvement from transferring a FlyWire-pretrained representation.

## G Population operating point and compute

**Population prior.** The balanced evaluation panels answer discrimination questions but do not represent deployment prevalence. When we reweight the Random test sets of Table 2 to FlyWire's 0.019% edge prior, the top 5% of ranked pairs retains $43.9\% \pm 2.0$ of positives at 0.165% precision (about $9\times$ enrichment) for Connectoformer, and $45.4\% \pm 1.8$ at 0.171% for the matched MLP (mean ± SD over 20 folds). The saved test sets are too small to estimate AP or rarer operating points reliably. This is a candidate-generation operating point for human review, not autonomous edge insertion.

**Model and inference cost.** In the runs of Table 2, the matched MLP has 191,553 parameters and Connectoformer + summary 1,246,060. On a fixed 4,096-pair benchmark on one A100, incremental cached-scoring allocation is 6.0 MB and 113.4 MB, respectively. Connectoformer's large one-time

peak comes from encoding 6,941 neurons; after those encodings are cached, scoring takes 0.072 ms per pair (approximately 13,900 pairs/s). The $O(K^2)$ cross-attention is paid again for every candidate pair over at most $K = 256$ retained tokens per neuron; only independent trunk encoding and partner-independent slots are reusable. Given these costs, a possible two-stage workflow would use the MLP or geometry for broad screening, then Connectoformer to rerank a shortlist of nearby candidate partners.

**Training compute.** Each run uses one A100- or H100-class GPU, with checkpointed resumes and a 1,000-epoch cap. The study behind Table 2 comprises 300 runs: ten models on 20 Random folds, five Region seeds, and five Patch seeds. A single Connectoformer run takes up to about a day. Across the project, cluster accounting records about 11,500 GPU-hours over roughly 5,500 jobs (January–October 2026), including development and runs not reported here. Fixed-checkpoint analyses (coordinate, feature-family, exact-type, localization, and prior-sensitivity audits) require no training and are much cheaper.

## H Extended related-work discussion

**Connectome resources.** Complementary *Drosophila* resources include the adult hemibrain [41], the FlyWire visual-system atlas [26], and the larval brain connectome [58]. Partial cortical reconstructions exist in mouse [29] and human [45]. Network-level analyses of the FlyWire graph itself have characterized motifs and brain-wide statistics [25]. FlyWire supplies our training and held-out-neuron panels, while the male adult CNS supplies the cross-individual test (Appendix F.1). The hemibrain, visual-system atlas, larval, and cortical resources provide context, not additional independent evaluation panels in this study.

**Peters' rule mechanism: appositions and pruning.** The Stepanyants neurogeometric framework [46, 48, 49] formalizes Peters' rule through axonal-dendritic appositions and shows that geometric overlap predicts cortical projection strength only up to a layer- and column-dependent scale factor. The Blue Brain algorithm [35] converts these appositions into synapses via a bouton-density-constrained pruning rule. Our arbor-overlap baselines on $32^3$ and $128^3$ grids are discretized analogues of the Stepanyants potential-synapse score; the super-linear connection-rate curve in Fig. 2(a) is consistent with the layer-dependent scale-factor finding.

**Point Affinity Transformer (PAT) contrast.** PAT [52] also uses cross-attention over per-point tokens to produce pairwise affinities, but differs from Connectoformer in three ways: (i) PAT's task is *within-neuron* point grouping (segmenting one neuron from a multi-neuron volume) whereas ours is *between-neuron* connectivity prediction; (ii) PAT consumes raw point coordinates whereas Connectoformer consumes branch-segment tokens with 18 morphometric features each; and (iii) PAT uses individual point-pair affinities as the prediction target whereas Connectoformer aggregates the full $K \times K$ similarity matrix into a single connectivity logit via a log-mean-exp pool, reserving the marginals for per-node synapse-localization (§5.4). Architecturally we follow the point-cloud and biological-sequence transformer lineage [18, 37, 61] with branch-segment features and learned tree-position embeddings.

## I Branch-segment vs. node tokenization: empirical comparison

We compare branch-segment tokens against the natural point-cloud alternative (skeleton points subsampled by farthest-point sampling, FPS) on three pre-model properties: token count, spatial coverage, and per-token discriminability. Numbers below are from $N{=}500$ FlyWire neurons.

**Token count.** Branch tokenization yields a median of 320 tokens per neuron (p25/p75 = 198/546, p99 = 3,785). Node-FPS at the original skeleton resolution yields a median of 3,022 points per neuron (p25/p75 = 1,782/5,858, p99 = 36,073). Within the encoder's sequence-length cap of $K{=}2{,}048$, branch tokenization fully captures 95.8% of neurons; node tokenization captures only 32.0% of neurons. Under an alternative truncation policy, 68% of node-token sequences would lose skeleton coverage; this is not the exclusion policy used by the headline branch-token model.

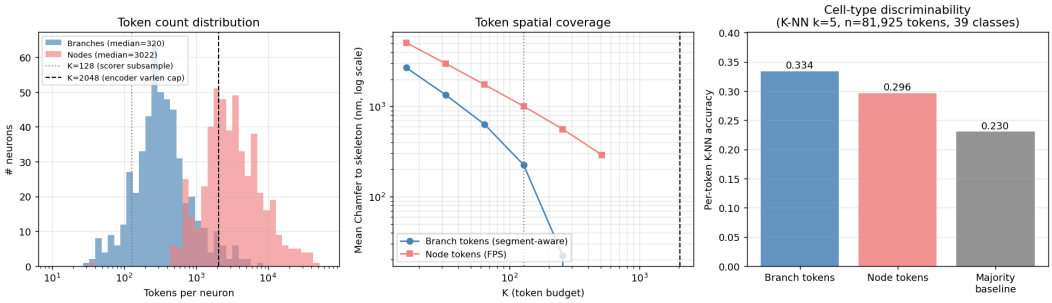



Figure 10: **Branch-segment vs. node tokenization.** *Left*: token count distribution across 500 neurons. Branches cluster around a median of 320 tokens/neuron; node counts span 1,800–6,000 with a long tail to 75,000. Vertical lines mark the illustrative scorer subsample ($K$=128, dotted) and the encoder sequence-length cap ($K$=2,048, dashed). The main scorer uses $K$=256. *Middle*: mean Chamfer distance from every skeleton point to the nearest selected token, sweeping the per-neuron token budget $K$. Branch-token coverage is segment-aware (a selected segment covers every point inside it). Branch tokens have lower coverage error at every $K$, reaching zero by $K$=512. *Right*: per-token K-NN classification accuracy on parent-neuron cell type using the same 8-D feature schema for both methods (relative position, local direction, normalized depth, normalized soma distance). Branch tokens carry more discriminative information per token despite occupying fewer slots.

**Spatial coverage.** At a fixed token budget $K$, we measure the mean distance from every skeleton point to its nearest selected token (segment-aware for branches: a selected segment covers all its constituent points at zero distance). At the illustrative budget $K$=128 (not the headline scorer's $K$=256 budget), branch tokens reach a median Chamfer of 225 nm vs. 1,004 nm for node-FPS (4.5× better). Branches saturate to zero by $K$=512; node-FPS still has 290 nm residual at that budget. The full sweep is in Figure 10.

**Per-token semantics.** Using a matched 8-D feature schema for both tokenizations (relative position, local direction, normalized tree depth, normalized soma distance), we train a $K$=5 nearest-neighbor classifier on individual tokens to predict their parent neuron's cell type. When we split training and test sets by neuron rather than by token, the classifier reaches 33.4% accuracy with branch tokens versus 29.6% with node tokens (against a 23.0% majority baseline, 39 cell-type classes, 81,925 tokens). The observed +3.8-point difference compares tokenization granularity under the same feature schema; it does not independently isolate all sampling or density effects.

**Reproducing this analysis.** The script `eval/analyze_tokenization.py` reproduces all three panels from raw SWC files in approximately fifteen minutes on a single CPU; no model checkpoint or GPU is required.

## J Auxiliary-loss ablation

These runs use the single-partition benchmark of Appendix L.1 (five optimization seeds, SE). Compare rows within this table, not with Table 2. One ablation supplies synapse density as an input; the main morphology-only model never receives this channel at inference.

Table 16: Loss-ablation on the random split. Mean ± standard error over 5 seeds. *Aux input* = synapse-density channel appended to token features; *Aux loss* = $\mathcal{L}_{\text{rank}} + \mathcal{L}_{\text{node}}$ on cross-attention marginals. Pair AUROC is overall test AUROC; Hard Neg is the soma-KNN hard-negative stratum; Loc AUROC is per-node synapse-localization AUROC averaged over pre/post sides.

| Variant | Aux input | Aux loss | Pair AUROC / Hard Neg | Loc AUROC |
|---|---|---|---|---|
| BCE only | — | — | 0.888±.002 / 0.854±.002 | 0.493±.014 |
| BCE + syn-derived input | yes | — | 0.888±.002 / 0.854±.003 | 0.492±.013 |
| Headline ($+\mathcal{L}_{\text{rank}} + \mathcal{L}_{\text{node}}$) | — | yes | 0.882±.003 / 0.845±.002 | 0.711±.011 |

Adding node-distribution and margin-ranking supervision raises the pair-conditioned localization AUROC by 0.218, with Overall/Hard AUROC changes of $-0.006/-0.009$. Localization supervi-

sion therefore incurs a small measured connectivity cost. Adding a synapse-density channel without these losses leaves localization AUROC near chance.

## K Visual depiction of the three train/test splits

We render each partition in dorsal soma-position space (XY projection of the FlyWire coordinates) for the 138,584 neurons with entries in both the coordinate and connection files. Train neurons are shown in blue, validation neurons in amber, and test neurons in red; for the Patch split, the 30 patch centers are overlaid as black star markers and the 75 $\mu$m-radius patches as dashed circles.

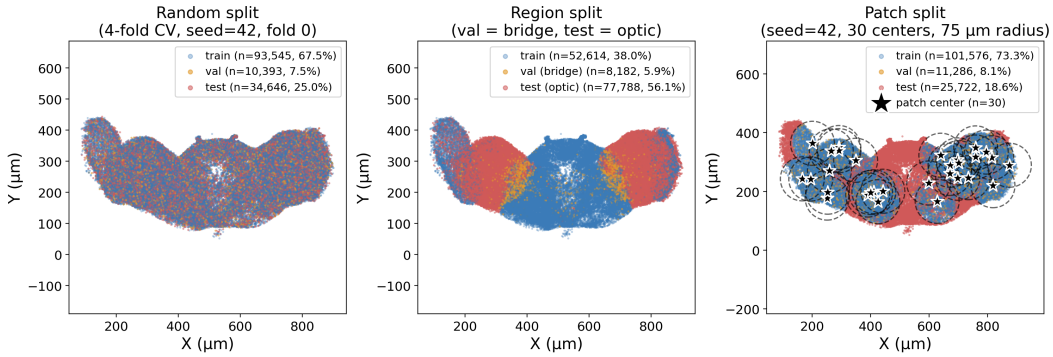



Figure 11: **The three splits at a glance.** *Random* (four-fold CV, seed 42, fold 0): a quarter of the neurons, spread across the brain, are held out for testing. *Region*: train on central and other superclasses (blue), select checkpoints on visual projection and centrifugal neurons at the optic-lobe boundary (amber), and test on the optic lobe (red). *Patch* (seed 42): training and validation neurons lie within 75 $\mu$m of 30 random soma centers; all other neurons are test.

## L Single-partition three-split benchmark

These tables report an earlier version of the benchmark: one 80/20 neuron partition for Random, five optimization seeds per split, Region trained on all non-optic neurons with no out-of-region validation set, Connectoformer checkpoints selected on test AUROC, and MLP rows using final-epoch weights after 300 epochs, with no checkpoint selection. We keep them for the additional metrics they report (AP, Overlap Hard, Directed, and Weak-edge AUROC); Table 7 reports the protocol used throughout the paper. Differences between the two should not be read as paired model changes.

### L.1 Random-split results

Table 17: **Random-split full results** (5-seed mean ± SE). Bold = best metric performance. -- = metric undefined for that arm.

| Method | AUROC | AP | Hard Neg | Ovlp HN | In-Type | Directed | Weak |
|---|---|---|---|---|---|---|---|
| Hemilineage | 0.53±.002 | 0.60±.002 | 0.41±.001 | 0.55±.001 | 0.13±.001 | 0.50±.000 | 0.50±.001 |
| Soma proximity | 0.62±.005 | 0.60±.005 | 0.04±.001 | 0.63±.002 | 0.04±.001 | 0.50±.000 | 0.49±.003 |
| Arbor overlap | **0.90±.002** | 0.87±.004 | 0.65±.001 | — | 0.53±.002 | 0.50±.000 | 0.62±.001 |
| MLP (token stats) | 0.83±.003 | 0.86±.005 | 0.79±.005 | 0.74±.006 | 0.81±.005 | 0.54±.001 | 0.57±.005 |
| MLP (summary) | 0.87±.002 | 0.90±.004 | 0.79±.002 | 0.78±.001 | 0.80±.002 | 0.56±.002 | 0.61±.003 |
| MLP (token stats + summary) | 0.87±.001 | 0.90±.002 | 0.79±.004 | 0.78±.005 | 0.80±.004 | 0.56±.002 | 0.60±.006 |
| Connectoformer (tokens only) | 0.84±.005 | 0.87±.003 | 0.81±.003 | 0.76±.003 | 0.83±.002 | 0.57±.005 | 0.62±.003 |
| Connectoformer + summary | 0.88±.003 | **0.90±.003** | **0.85±.002** | **0.80±.002** | **0.86±.001** | **0.61±.005** | **0.64±.003** |

### L.2 Region-split results

The region split trains on the diverse central + sensory + ascending + descending + motor + endocrine super-classes (∼61K neurons) and tests on the held-out optic-lobe super-class (∼78K neurons): a

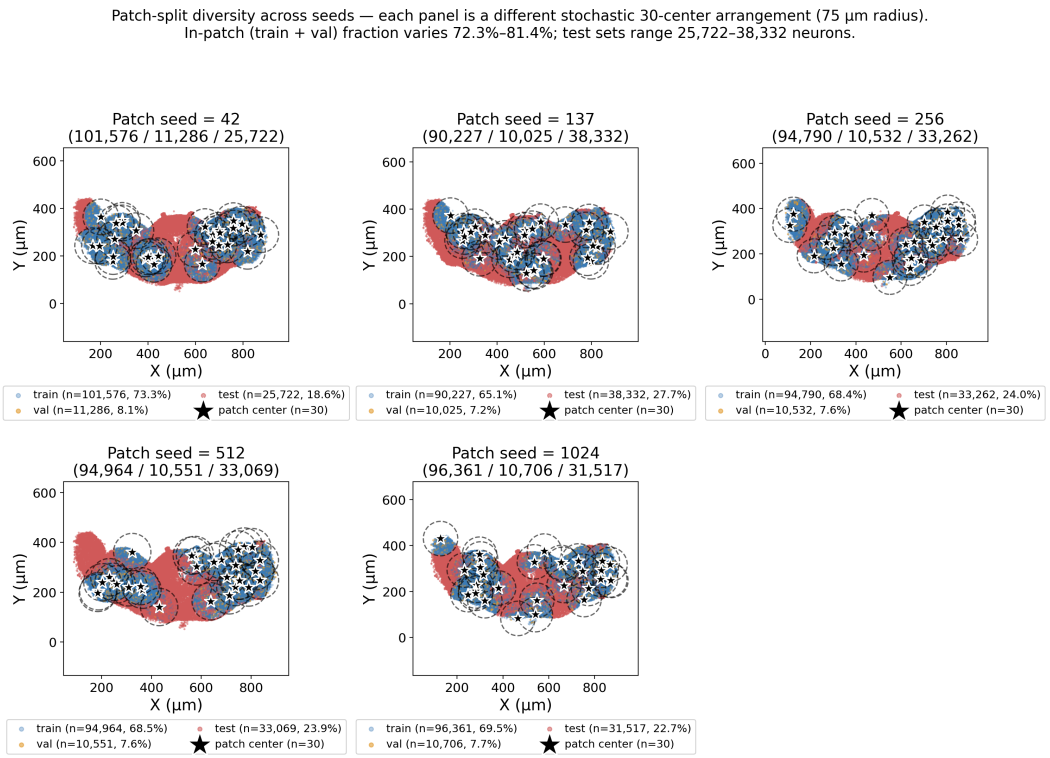



Figure 12: **Patch-split partition variance across the 5 seeds used in the paper.** Each panel shows a different stochastic 30-center arrangement (seeds $42, 137, 256, 512, 1024$; $75\,\mu$m radius). The in-patch (training plus validation) fraction varies 72.3–81.4%, producing test sets of 25,722–38,332 neurons, a roughly $1.5\times$ spread that contributes to the larger SE on Patch-split metrics in the main paper.

training distribution that spans the morphological diversity of the central nervous system, evaluated on a homogeneous columnar visual-processing region the model never saw at training time.

Table 18: **Region-split full results** (5-seed mean ± SE). Bold = best metric performance.

| Method | AUROC | AP | Hard Neg | Ovlp HN | In-Type | Directed | Weak |
|---|---|---|---|---|---|---|---|
| Hemilineage | 0.53±.004 | 0.52±.002 | 0.49±.003 | 0.56±.003 | 0.24±.003 | 0.50±.000 | 0.61±.003 |
| Soma proximity | 0.48±.001 | 0.44±.000 | 0.06±.002 | 0.63±.002 | 0.06±.002 | 0.50±.000 | 0.63±.002 |
| Arbor overlap | **0.81±.002** | 0.73±.004 | 0.61±.002 | **0.90±.001** | 0.55±.002 | 0.50±.000 | 0.65±.003 |
| MLP (token stats) | 0.73±.002 | 0.75±.002 | 0.73±.002 | 0.67±.002 | 0.72±.002 | 0.54±.002 | 0.37±.004 |
| MLP (summary) | 0.76±.004 | 0.78±.004 | 0.76±.004 | 0.69±.001 | 0.75±.004 | 0.54±.002 | 0.37±.001 |
| MLP (token stats + summary) | 0.76±.004 | 0.78±.003 | 0.76±.002 | 0.69±.001 | 0.75±.002 | 0.54±.002 | 0.38±.003 |
| Connectoformer (tokens only) | 0.76±.003 | **0.79±.002** | 0.77±.001 | 0.68±.001 | 0.76±.001 | 0.53±.002 | 0.44±.003 |
| Connectoformer + summary | 0.77±.003 | 0.79±.002 | **0.77±.002** | 0.69±.003 | **0.76±.002** | **0.54±.004** | 0.44±.006 |

## L.3 Patch-split results

The Patch split trains on neurons within random $75\,\mu$m soma-neighborhoods and tests on the rest (stochastic per-seed): different seeds produce different patch positions, so the SE bars include partition variance jointly with model variance. This complements the random and region splits with a graded out-of-distribution test that does not depend on a single fixed anatomical partition.

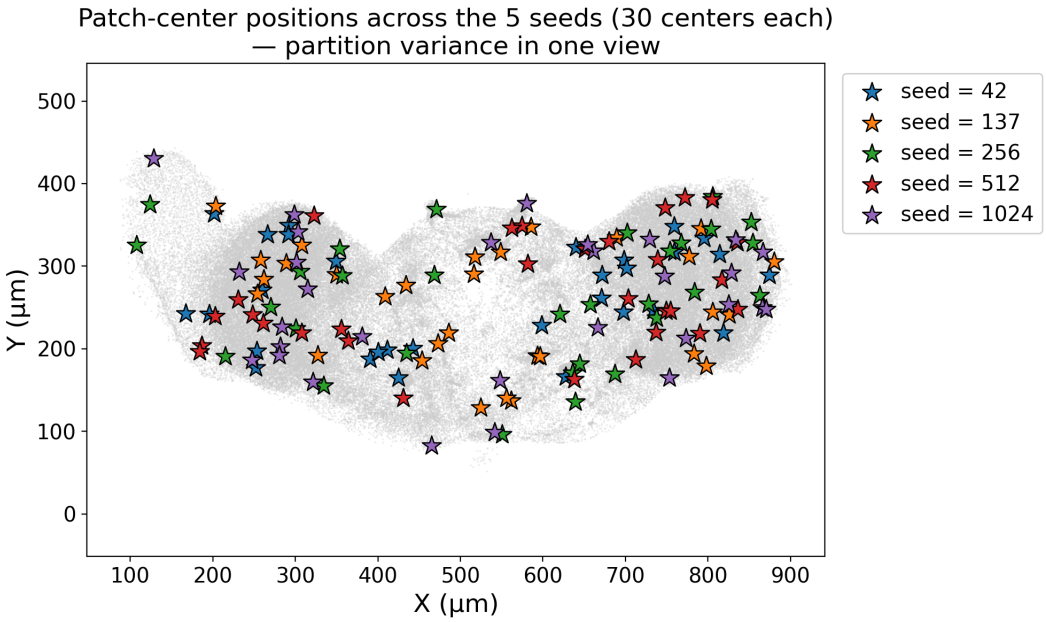


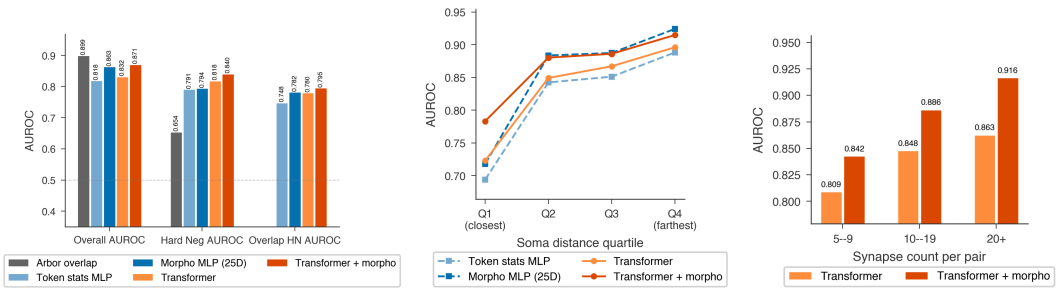

Figure 13: **Patch centers across all 5 seeds in one view.** The full neuron population is rendered as a faint gray background; the 30 centers per seed are overlaid as differently colored stars. Center positions visibly shift between seeds, which is the source of partition variance the Patch split is designed to capture.

Table 19: **Patch-split full results** (5-seed mean ± SE). Bold = best metric performance.

| Method | AUROC | AP | Hard Neg | Ovlp HN | In-Type | Directed | Weak |
|---|---|---|---|---|---|---|---|
| Hemilineage | 0.48±.008 | 0.49±.004 | 0.37±.013 | 0.51±.010 | 0.11±.012 | 0.50±.000 | 0.49±.012 |
| Soma proximity | 0.44±.005 | 0.43±.003 | 0.06±.006 | 0.46±.007 | 0.05±.005 | 0.50±.000 | 0.48±.015 |
| Arbor overlap | **0.85±.001** | 0.77±.004 | 0.69±.006 | **0.84±.011** | 0.54±.013 | 0.50±.000 | 0.66±.007 |
| MLP (token stats) | 0.80±.007 | 0.79±.007 | 0.77±.010 | 0.71±.011 | 0.76±.010 | 0.53±.005 | 0.56±.022 |
| MLP (summary) | 0.84±.007 | 0.82±.011 | 0.79±.012 | 0.71±.022 | 0.77±.013 | 0.53±.005 | 0.59±.012 |
| MLP (token stats + summary) | 0.83±.005 | 0.82±.006 | 0.79±.008 | 0.71±.016 | 0.77±.009 | 0.52±.007 | 0.58±.011 |
| Connectoformer (tokens only) | 0.81±.005 | 0.79±.006 | 0.79±.003 | 0.74±.005 | 0.79±.006 | 0.54±.006 | 0.58±.013 |
| Connectoformer + summary | 0.85±.007 | **0.82±.007** | **0.81±.006** | 0.76±.006 | **0.80±.007** | **0.56±.006** | 0.57±.011 |

## M Stratified breakdowns by distance and synapse count

Figure 14 shows how discrimination varies with the evaluation stratum, soma distance, and connection strength on the single-partition benchmark of Appendix L.1; its values are comparable only within that benchmark.




(a) Overall and hard-negative AUROC across models.

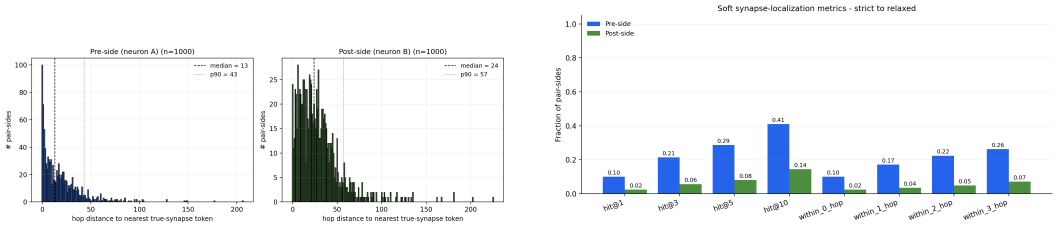
(b) AUROC by soma-distance quartile.

(c) AUROC by synapse-count bin.

Figure 14: **Stratified connectivity results (single-partition benchmark). (a)** Arbor overlap scores highest overall in this panel, while summary-conditioned Connectoformer scores highest on its two hard strata. **(b)** Distance stratification distinguishes close-range partner discrimination from rejection of distant negatives. **(c)** More strongly connected pairs are more reliably predicted from morphology in this evaluation.

## N Token retrieval plots

**Token retrieval.** The 1,000-pair diagnostics of Table 13 measure the tree-hop distance from the top-ranked token to the nearest true synaptic token, and the probability of retrieving a true synaptic token among the top $k$ full-sequence candidates. Hit@1 is $10.0\%/2.4\%$ and Hit@10 is $41.0\%/14.4\%$ on the pre/post sides. Median distances are $13/24$ token-tree hops (90th percentiles $43/57$); these are topological distances, not physical distances in nanometers. The pre/post asymmetry is visible under both diagnostics; its cause is not identified by these plots alone.




(a) Tree-hop distance from the top-ranked token to the nearest true synaptic node ($n = 1{,}000$ pairs).

(b) Hit@$k$ and retrieval within increasing tree-hop tolerances.

Figure 15: **Token-level localization on the FlyWire test set,** on the full sampled population of Table 13 (including above-training-cap neurons). The presynaptic side has tighter tree-hop distances and higher retrieval rates. Relaxing $k$ or the allowed tree-hop distance increases coverage of true synaptic nodes. Strict retrieval remains the principal estimate of precise token selection.

## O Shuffled-morphology null tests

The neuron-disjoint train/test split (§3) already prevents the model from memorizing per-neuron labels; test pairs use neurons the model has never seen during training, on either side. To test whether the baselines depend on the correct morphology-to-neuron mapping, we also shuffled that mapping before training: each neuron identifier is paired with another randomly chosen neuron's skeleton features, breaking any link between morphology and the labeled connectivity graph.

If a model still achieves above-chance AUROC under this perturbation, it is exploiting label-side or pair-level statistical shortcuts rather than the morphology itself. The tested non-identity baselines drop near chance ($\approx 0.50$) on both splits. This is a supportive control, not a proof that all identity or pair-level shortcuts have been eliminated.

Table 20: Shuffled-morphology null tests. Skeleton-neuron assignments are randomized prior to training, then the model is trained and evaluated on the same downstream task. Values near 0.50 support a dependence on the morphology-to-neuron mapping in these tested baselines, without eliminating every shortcut.

| Method | Input | Random AUROC | Region AUROC |
|---|---|---|---|
| Shuffled MLP (token stats) | 54D pooled | 0.517 | 0.509 |
| Shuffled MLP (summary) | 25D hand-crafted | 0.484 | 0.524 |

# P Per-branch feature attribution

The attribution plot summarizes mean absolute gradients of the pair score with respect to token features on 100 test pairs. It describes which inputs affect this trained scorer locally; retrained component controls are reported separately in Appendix D.3.

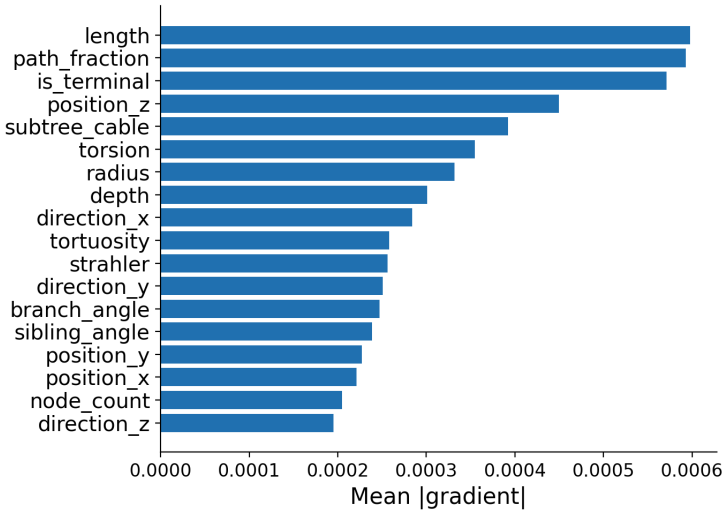



Figure 16: **Per-branch gradient attribution** on the cross-attention scorer ($n = 100$ test pairs). Branch length, path fraction, terminality, position, and subtree cable contribute to the sensitivity profile. The diffuse importance pattern describes reliance on several coupled features, without identifying independent biological effects.

# Q C. elegans data provenance

## Q.1 Dataset lineage and provenance

We derive all *C. elegans* data from a single source: the `dwitvliet/nature2021` GitHub repository (`https://github.com/dwitvliet/nature2021`) accompanying Witvliet et al. [59]. We deliberately avoid mixing data from multiple processing pipelines (e.g., the BossDB-hosted VAST synapse exports) because the two annotation passes use different coordinate frames; combining them would require registration. We avoid this mismatch by sourcing skeletons and synapses from the same CATMAID export.

**Vertex-linked synapse positions.** Witvliet's synapse table records each synapse by a pre-synaptic *treenode* ID (`pre_tid`) and a list of post-synaptic treenode IDs (`post_tid`). These IDs index directly into the matching skeleton's vertex set, so no spatial projection is needed to attach a synapse to a skeleton vertex. We verified the linkage on all 3,674 polyadic synapses: $3{,}674/3{,}674$ pre-vertex IDs map to a valid skeleton vertex; $7{,}909/7{,}992$ post-vertex IDs do (the 83 misses are synapses whose post neurons lack a curated skeleton in this dataset and are therefore dropped from `synapse_locations.csv`).

Table 21: C. elegans data lineage. All files derive deterministically from two raw downloads and the `eval/celegans/witvliet_dataset.py` converter.

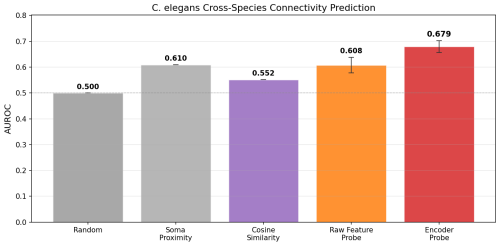
| File | Source | Transformation |
|---|---|---|
| `Dataset8_skeletons.json` | dwitvliet/nature2021 GitHub | byte-identical download |
| `Dataset8_synapses.json` | dwitvliet/nature2021 GitHub | byte-identical download |
| `swc/⟨name⟩.swc` (216) | dwitvliet (via skeletons JSON) | 7-col SWC from `coords` + `arbor` |
| `coordinates.csv` (216) | dwitvliet (via skeletons JSON) | soma = `coords[starts[0]]` |
| `neuron_names.csv` (216) | dwitvliet (via skeletons JSON) | sorted-alphabetical → 0-indexed IDs |
| `synapse_locations.csv` (7,909) | dwitvliet (via synapses + skeletons JSON) | polyadic → dyadic, position from `skel[pre][coords][pre_tid]` |
| `connections.csv` (2,155) | dwitvliet (via synapse_locations) | deduplicated $(pre, post)$ pairs |

**Polyadic-to-dyadic expansion.** Each *C. elegans* synapse is anatomically a single pre-synaptic active zone with one or more post-synaptic densities; in Witvliet's JSON this appears as one entry per active zone with a `post` list of length $N$. We expand each entry into $N$ dyadic rows, all sharing the same $(x, y, z)$ (= `skel[pre][coords][pre_tid]`). Sharing the active-zone location preserves the polyadic synapse's single presynaptic site and avoids spurious spatial separation between its dyadic rows.

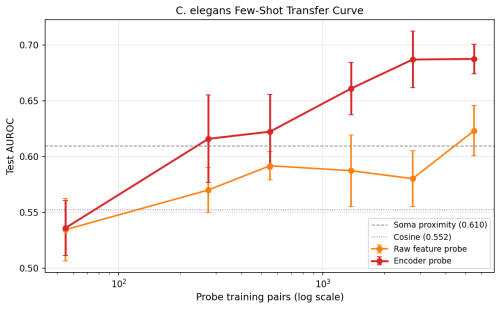
**Cross-validation against VAST.** For provenance verification we compared our Witvliet-derived `connections.csv` against the older VAST CSV exports on the same dataset: 2,155 pairs in both vs. 2,156 in VAST, 2,155/2,155 shared, only 1 pair appearing in VAST but not Witvliet. The two pipelines agree on connectivity to within $\sim 0.05\%$ at the pair level; we use Witvliet exclusively for the coordinate-frame consistency reasons above.

### Q.2 Single-panel cross-species probe figures

Figures 17 and 19 show an earlier, single-panel version of the probe and its training-pair sweep, in which the pretrained encoder reached AUROC 0.679 versus 0.608 for raw morphology. That advantage does not hold up under the repeated probes of Table 15, which use matched partitions and explicit label budgets and are the basis for our conclusion about pretraining.




(a) AUROC comparison, including the five-fold raw-feature and pretrained-encoder probes.

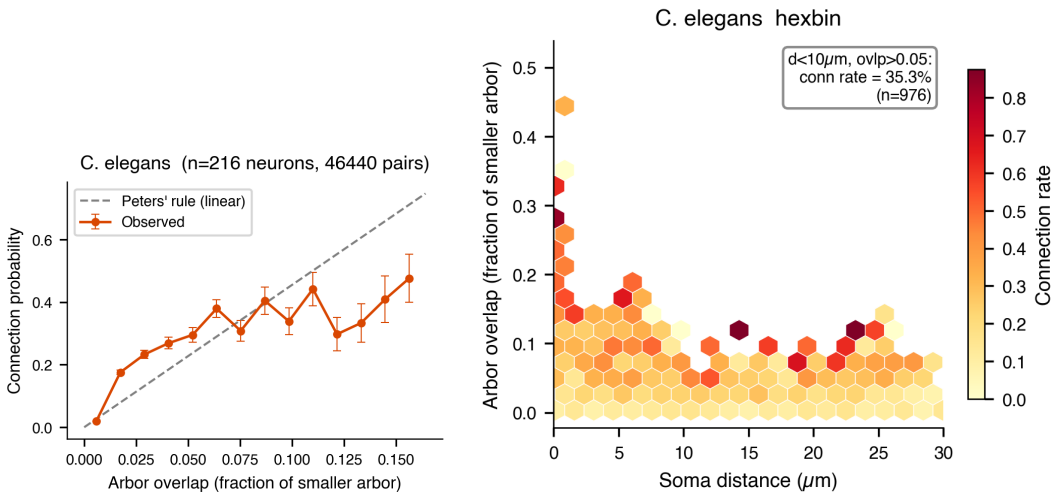



(b) Few-shot sweep over labeled training pairs; bars show standard errors over neuron-disjoint folds.

Figure 17: **Single-panel cross-species probes.** *C. elegans* adult connectome [59, dataset 8]: 216 neurons and 2,155 directed connections. The higher pretrained-probe means in this panel do not hold up under the repeated protocol of Table 15.

### Q.3 Geometric connectivity diagnostics in C. elegans

The geometric diagnostic enumerates all $216 \times 215 = 46{,}440$ directed pairs and measures arbor co-occupancy on a $96^3$ grid. The curves describe how geometric opportunity relates to observed connectivity in this smaller, denser connectome.



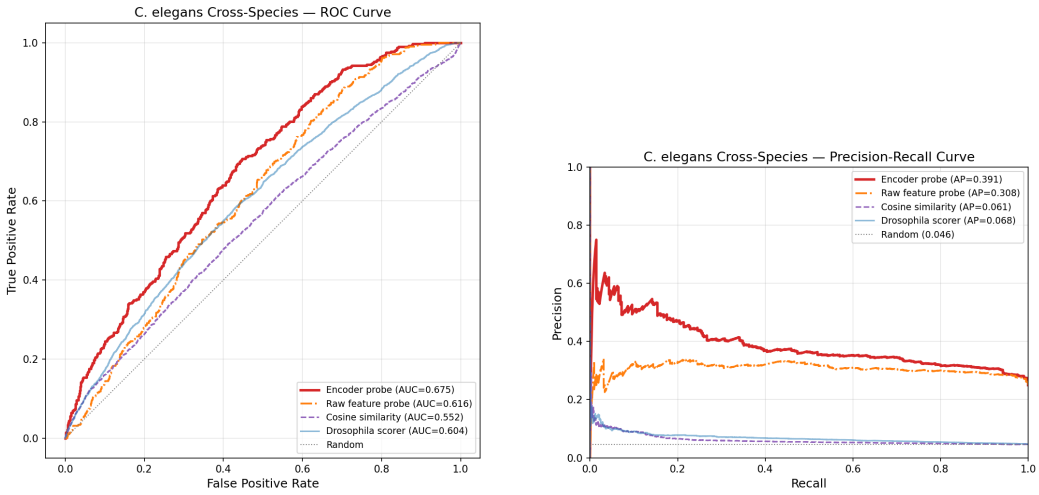

(a) Connection probability versus arbor overlap, with linear Peters-rule reference fits.

(b) Distance versus overlap, colored by connection rate for pairs closer than 30 $\mu$m.

Figure 18: **Peters-rule diagnostics in C. elegans.** 216 neurons, 46,440 directed pairs, and $96^3$ voxelization (approximately 357 nm voxels). The overall connection rate is 4.6%; among overlap-bearing pairs it is 19.6%. The relationship between overlap and connection probability departs from a single proportional rule.

## Q.4 ROC and precision-recall curves

These curves come from the single-panel probe of Fig. 17. They describe discrimination across thresholds and the effect of class imbalance; they do not substitute for the repeated comparisons in Table 15.




(a) Receiver-operating-characteristic curves.

(b) Precision-recall curves at a 4.6% positive rate.

Figure 19: **C. elegans discrimination curves.** The probe curves use the single-panel protocol of Fig. 17; the repeated pretraining comparisons and confidence intervals are in Appendix F.2.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim morphology-based connectivity prediction beyond geometric opportunity, hard-case gains under matched inputs, group and cross-individual evidence within *Drosophila*, and coarse auxiliary synapse-site ranking. Tables 2, 3, and 13 support those claims; §6 bounds cross-species and precise-localization interpretations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The introduction states that the analysis is predictive rather than causal. §6 covers the 4.1% long-sequence exclusion, input matching without parameter or auxiliary-supervision matching, the bridge-neuron proxy used for Region checkpoint selection, group-dependent transfer, sampling dependence in fixed panels, limited cross-species support, weak signal within fixed cell types, and coarse localization. Appendix G reports the repeated $O(K^2)$ candidate-pair scoring cost.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents an empirical methodology and architecture; no formal theorems or proofs are claimed.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §4 and Appendix B specify tokenization, shared encoding, four-slot pooled scoring, cross-attention, and auxiliary losses. Table 6 distinguishes the repeated Random, Region, Patch, and group-held-out units; Appendix C defines pairing and uncertainty. Code to reproduce training and evaluation is available at `https://github.com/sgoe1/connectoformer`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code for preprocessing, training, baselines, and the evaluation analyses, including the male CNS and *C. elegans* transfer analyses, is available at `https://github.com/sgoe1/connectoformer` under the MIT license. Its README gives installation instructions, the configuration file and command for each reported experiment, and a no-data smoke test. The FlyWire, male CNS, and Witvliet *C. elegans* data are publicly available; the repository includes scripts to download (male CNS, *C. elegans*) or preprocess (FlyWire) them and to build every derived cache. Trained checkpoints and preprocessed caches are not distributed. The architecture, training protocol, splits, hyperparameters, and ablations are also specified in §4, §4, and Appendices B and D.3.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: §4 and Appendix B give optimizer, learning rate, weight decay, warmup, epoch cap, early stopping, pair sampling, and auxiliary weights. §3 and Table 6 distinguish split and replicate policies; Appendix D.3 distinguishes retraining from fixed-checkpoint masking.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Captions state whether values are SE over five seed-level units or SD across held-out folds. Random paired inference averages four folds within each of five seeds; Region and Patch use five seeds. We report two-sided $t$-based 95% CIs and Holm correction over seven declared contrasts by split/endpoint. Group-held-out intervals use five folds. Fixed-panel intervals are conditional on trained models, and shared-neuron dependence is explicitly discussed. Fixed-prior paired tests use retrained, validation-selected runs rather than being inferred from SEs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix G reports the GPU types, the size of the main study, per-run training time, total project GPU-hours from cluster accounting, and measured cached-inference costs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses publicly released connectome datasets (FlyWire, the male adult *Drosophila* CNS, and the Witvliet *C. elegans* adult connectome) with no human subjects, no sensitive personal data, and no foreseeable dual-use risks.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is basic research on morphology-based connectivity prediction in animal connectomes. Its proposed applications are prioritizing candidate connections for experimental testing and human connectomics review, not autonomous edge insertion or inference about human subjects. No direct societal deployment is proposed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released code operates on animal-neuron morphology and contains no sensitive personal data. We identify no high-risk deployment that would require controlled model access; prediction and transfer limitations are documented.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets are cited and used under their stated terms: FlyWire v783 [10, 42] (CC BY-NC 4.0 under the FlyWire data guidelines), the male adult *Drosophila* CNS [2] (CC BY 4.0), and the Witvliet et al. *C. elegans* data [59] (no license stated; used for non-commercial research only). Provenance is in Appendices A and Q.1. We release code only; no data or derived caches are redistributed.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset is the Connectoformer code (`https://github.com/sgoe1/connectoformer`), released under the MIT license. Its README documents installation, data preparation, training, and the configuration and script for each experiment, and the repository includes a test suite with a no-data training smoke test. No new datasets or trained checkpoints are released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a core model component or an original scientific method in this work. AI assistance was used in coding, analysis tooling, experiment orchestration, and manuscript editing; scientific results are grounded in recorded experiments and author verification, not generated claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.