LEARNING FROM THE OBSERVATION GRAPH OF INCOMPLETE CONNECTOMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Connectome reconstruction records synaptic relations before every observed segment has a resolved neuronal identity. Restricting inference to traced neurons can therefore discard evidence that is already available. We formulate learning on this evolving *observation graph* and distinguish the value of preserving unresolved relations from the additional value of learning to combine them. Our leave-fragment-out supervision samples count-conserving queries from traced source neurons and subtracts each query from its positive owner before feature construction, removing direct query self-overlap. Models trained in MaleCNS are frozen and evaluated on recorded MANC revisions: 557 donor segments in 35 components, each ranked against all 23,200 earlier traced candidates. The segment-aware neural scorer achieves component-macro Hit@20 of 0.5399, compared with 0.4866 for its retained-only counterpart and 0.4666 for fixed cosine. Gains depend on the inspection budget: Hit@1, Hit@5, and mean reciprocal rank decrease. A separate masked-adjacency task improves from 0.3800 to 0.4708 average precision when unresolved context is retained. Correspondence interventions and support-stratified analyses show where that context contributes and where a fixed rule remains sufficient. Together, these results support observation preservation as a useful design choice and identify a conditional benefit of source learning for structural retrieval. The evaluated outcome is a recorded revision destination; biological ownership and proofreading productivity remain separate endpoints.

1 INTRODUCTION

A connectome is usually represented as a graph whose nodes are neurons and whose edges are synapses. In an evolving reconstruction, however, observed synapses may connect segments that have not yet been assigned to complete neurons. This gap between observed connectivity and resolved identity persists even as reconstructions reach whole adult brains (Dorkenwald et al., 2024). Infrastructure such as CAVE tracks changing assignments while preserving access to annotations through time (Dorkenwald et al., 2025). The learning question is immediate: *should an unresolved neuronal identity prevent an observed relation from contributing to inference?*

Existing reconstruction systems establish the value of evidence at several scales. NEURD extracts morphological structure to support automated correction (Celii et al., 2025); Autoproof learns from proofreading decisions to prioritize work and attach fragments (Huang et al., 2025). Recent benchmarks extend the evaluation of split and merge decisions across image and mesh inputs (Brown et al., 2026). These advances motivate a complementary question about representation: how much useful information is lost when a learning pipeline keeps only relations among already traced bodies? Resolving this question requires separating the value of the additional observations from the advantage of a learned scoring rule.

We study this distinction through two tasks. The primary task retrieves a donor segment's later recorded destination from the complete set of earlier traced candidates. MaleCNS provides source supervision (Berg et al., 2026), and MANC provides a separate target individual and version-derived outcomes (Takemura et al., 2024). A supporting masked-adjacency task tests whether unresolved relations carry predictive information when the candidate relation itself is withheld. In both cases, the input is a directed connectivity profile; image and mesh evidence are outside the evaluated input space.

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

The main methodological obstacle is a supervision shortcut. A query sampled from a known owner can be retrieved through direct self-overlap if its counts remain inside the positive candidate. We remove the query before computing any candidate-dependent statistic. Models fitted to these residualized source tasks are then frozen before target scoring. On the target, features come exclusively from the earlier release, and evaluation averages first within each recorded revision component. These choices control distinct sources of optimism: answer-containing candidates, later-assignment leakage, and unequal weighting of shared edits.

The resulting evidence supports a specific claim. A small neural scorer using retained and unresolved profiles improves twenty-candidate coverage over both its matched retained-only arm and fixed cosine, while reducing accuracy at shorter budgets. Support-stratified and correspondence controls localize this gain. Our contributions are (i) an observation-graph formulation that separates information preservation from learned combination; (ii) count-conserving source supervision with an exact self-overlap analysis; and (iii) full-candidate transfer evaluation with explicit ties, component weighting, and operating-point trade-offs. Together, they provide a way to study learning under incomplete identity without assuming that a more complex graph encoder is the source of the gain.

2 Observation graph, supervision, and endpoint

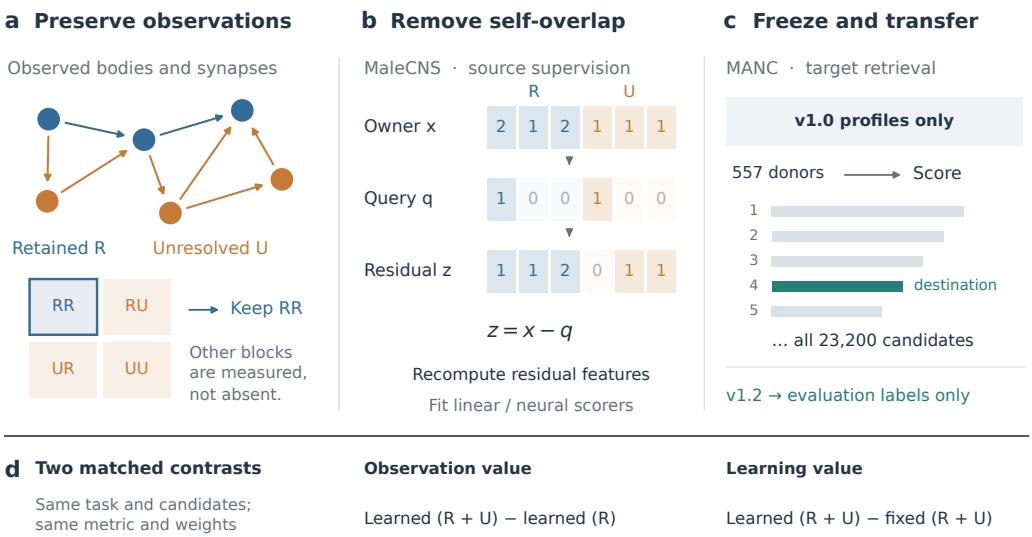

Figure 1: **Learning before neuronal identity is complete. a,** Retaining only traced bodies (R) removes measured relations involving unresolved segments (U). **b,** A synthetic source query is subtracted from its owner before features are computed; the illustrated count vectors are schematic. **c,** Frozen source models rank all earlier traced target candidates. Later assignments supply evaluation labels only. **d,** Matched observation and learning contrasts answer separate questions. Colors consistently denote R (blue), U (orange), and the recorded destination (teal).

2.1 Observation space and two estimands

Let $W^{(t)} \in \mathbb{N}_0^{|V_t|\times|V_t|}$ count confidence-filtered directed synapses between observed body IDs at release t. Partition $V_t = R_t \dot\cup U_t$ into retained traced bodies and remaining observed segments. Here, unresolved is an operational designation for bodies outside the status-`Traced` selection, rather than a certified biological fragment category. With coordinate selector P_R,

$$W_{RR} = P_R W P_R^\top, \qquad W = \begin{pmatrix} W_{RR} & W_{RU} \\ W_{UR} & W_{UU} \end{pmatrix}. \tag{1}$$

The omitted cross-blocks contain observed incidence, not hypothetical missing edges. For channel $H \in \{R, U\}$, define outgoing and incoming profiles $x_i^{H,\mathrm{out}} = (W_{ij})_{j\in H}$ and $x_i^{H,\mathrm{in}} = (W_{ji})_{j\in H}$.

Under the MaleCNS status-`Traced` selection, the retained graph contains 42.1% of outgoing but 95.1% of incoming observed synapse-count mass. These are representation statistics, not biological sensitivity or a prediction effect.

Let f_R and f_{R+U} be matched learned scorers, and let h_{R+U} be a fixed rule using the full observation. Holding the metric M_K, candidates, and evaluation mask constant, we distinguish

$$\Delta_K^{\text{obs}} = M_K(f_{R+U}) - M_K(f_R), \qquad \Delta_K^{\text{learn}} = M_K(f_{R+U}) - M_K(h_{R+U}). \tag{2}$$

The first contrast measures the effect of access to the unresolved channel within a scorer family. The second asks whether learned combination improves on a specified fixed rule after preserving the same information. Neither contrast, by itself, establishes superiority over every architecture or fixed rule.

2.2 Recorded integration as conditional full-candidate retrieval

MANC v1.0 and v1.2 synapse rows are aligned by exact pre/post coordinates. Six coordinate-inconsistent rows are excluded. Changed assignments induce 35 qualifying many-to-one components whose same-ID destination was already traced in v1.0. Component g has donors D_g and recorded destination $y_g \in R_{\text{old}}$, with $\sum_g |D_g| = 557$. Every donor is ranked against the same $\mathcal{C} = R_{\text{old}}$, $|\mathcal{C}| = 23{,}200$, using only old-release profiles. Later assignments define outcomes and never enter target features. The endpoint is therefore conditional on recorded convergence and destination coverage; it does not detect every possible reconstruction edit.

3 Leave-fragment-out source learning

3.1 Count-conserving fragments remove direct self-overlap

Concatenate the four direction/channel profiles of source owner o as $x_o \in \mathbb{N}_0^D$, with $n_o = \sum_j x_{oj}$. For $m \in \{1, 2, 4, 8\}$, sample endpoint counts without replacement,

$$\Pr(q \mid x_o, m) = \frac{\prod_j \binom{x_{oj}}{q_j}}{\binom{n_o}{m}}, \quad 0 \le q_j \le x_{oj}, \quad \sum_j q_j = m. \tag{3}$$

For this query, set $z_o = x_o - q$ and $z_c = x_c$ for $c \neq o$. Then $q + z_o = x_o$, whereas leaving the owner unchanged gives

$$q^\top x_o = q^\top (x_o - q) + \|q\|_2^2, \qquad m \le \|q\|_2^2 \le m^2. \tag{4}$$

The final term is direct query self-overlap. All candidate-dependent norms, strengths, binary supports, and resource-allocation document frequencies are recomputed after subtraction. Appendix A proves the corresponding finite-population expectation and shows that binary profiles have zero residual same-coordinate dot product.

3.2 Features, objective, and freezing

Each retained/unresolved channel contributes outgoing, incoming, and joint cosine; binary Jaccard; resource-allocation overlap; and four log strengths, giving 18 standardized features. We compare a linear scorer with a ReLU MLP $18 \to 32 \to 16 \to 1$ (1,153 trainable parameters). The off arm uses the identical feature pipeline but zeros the standardized unresolved channel.

The source universe contains 4,096 candidates and 256 owners, split owner-disjoint into 192 fitting and 64 validation owners. Four fixed fragments per owner yield 768 fitting and 256 validation tasks. Each task uses the same sampled list in both arms: one positive plus 96 negatives formed from hard retained/unresolved candidates and deterministic uniform filling. With V_o the four fragments of owner o,

$$\widehat{\mathcal{L}}(\theta) = \frac{1}{|O_{\text{fit}}|} \sum_{o \in O_{\text{fit}}} \frac{1}{|V_o|} \sum_{q \in V_o} \left[-s_\theta(q, o) + \log \sum_{c \in L_{o,q}} e^{s_\theta(q,c)} \right]. \tag{5}$$

The loss applies listwise normalization over sampled candidates (Cao et al., 2007). Linear/neural, on/off models are trained for 200 Adam updates with seeds 17, 29, and 43 (Kingma and Ba, 2015).

All twelve final checkpoints and source normalization moments are frozen before target scoring. This is sampled-list discrimination, not an unbiased full-candidate likelihood or a calibrated ownership probability.

3.3 Exact ties and component-macro evaluation

For destination score s_y, let a candidates score strictly higher and let the tie block containing y have size b. Uniform tie integration gives

$$\mathrm{Hit}_K = \frac{\min\{b, \max(0, K-a)\}}{b}, \qquad \mathrm{RR} = \frac{1}{b}\sum_{r=a+1}^{a+b}\frac{1}{r}. \tag{6}$$

Donors in one revision component share a destination. We therefore report

$$\widehat{M}_K = \frac{1}{35}\sum_{g=1}^{35}\frac{1}{|D_g|}\sum_{d\in D_g}\mathrm{Hit}_K(d, y_g), \tag{7}$$

with the analogous aggregation for reciprocal rank. This assigns equal weight to each recorded component. Three-seed means summarize optimization variation; the seeds do not provide biological replication.

4 Source learning improves coverage at a wider inspection budget

4.1 Full-candidate transfer reveals a budget trade-off

The segment-aware neural scorer reaches component-macro Hit@20 of 0.5399 on recorded MANC revisions (Table 1). The observation contrast is $\Delta_{20}^{\mathrm{obs}} = 0.0533$ relative to the matched neural R arm; the learning contrast is $\Delta_{20}^{\mathrm{learn}} = 0.0733$ relative to fixed joint cosine. All three neural seeds have positive contrasts. Linear scoring also benefits from the additional channel at this budget, increasing from 0.4840 to 0.5035.

Table 1: **Recorded-revision retrieval.** All methods rank 23,200 candidates for each of 557 donors; metrics average over 35 revision components. Learned scores average three source-training seeds. Bold denotes the column maximum. The primary budget is shaded, and all shorter-budget metrics are retained.

	Observation	Destination coverage ↑			Rank quality ↑
Scorer	channels	Hit@1	Hit@5	Hit@20	MRR
Fixed joint cosine	$R+U$	**0.1415**	0.3069	0.4666	**0.2200**
Linear	R	0.1088	**0.3289**	0.4840	0.2053
	$R+U$	0.0851	0.2985	0.5035	0.1762
Neural	R	0.1075	0.3197	0.4866	0.2036
	$R+U$	0.0840	0.2836	**0.5399**	0.1756

This benefit comes with a change in rank allocation. Neural $R+U$ falls below neural R at Hit@1, Hit@5, and MRR; fixed cosine has the highest Hit@1 and MRR among the displayed methods. Figure 2 makes the budget dependence explicit. The model improves the chance of including the recorded destination in a twenty-candidate list, while offering a weaker ordering for very short inspection budgets. Thus, the operating budget is part of the result rather than a detail to choose after deployment.

4.2 Correspondence contributes beyond candidate strength

Shuffling the unresolved relation features across candidates reduces neural Hit@20 from 0.5399 to 0.4907–0.4933 across three permutations. Setting these features to zero gives 0.4964. Because query

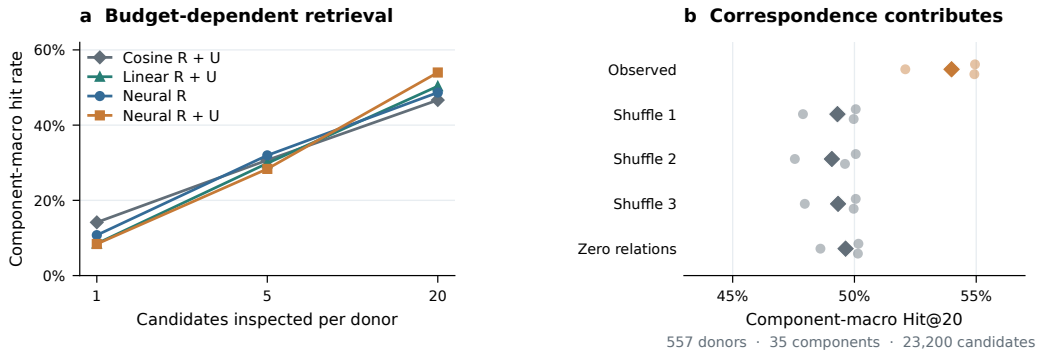

Figure 2: **Coverage gains depend on budget and relation correspondence. a,** Full-candidate retrieval at three budgets. **b,** Frozen neural $R + U$ under candidate-wise shuffling or zeroing of the five unresolved relation features; strengths remain fixed. Large markers show means and small dots show source-training seeds. The seeds measure fit variation, not uncertainty across biological specimens.

and candidate strengths remain fixed, this sensitivity identifies a contribution from candidate-specific relation correspondence at $K = 20$. Some shorter-budget metrics improve under intervention, further limiting the conclusion to the reported operating point. Complete linear and neural intervention results appear in Appendix D.

The gain also extends across multiple revision components. Neural $R+U$ is better, worse, or tied with neural R in 18, 4, and 13 components, respectively. Removing the five largest positive component differences leaves a mean improvement of 0.0172. This removal analysis establishes breadth within the observed cohort; it does not estimate a population effect from 35 independent biological samples.

4.3 Available support determines the useful comparison

Table 2: **Where does the added context help?** Donor-average Hit@20, grouped by available input support. These rows partition the 557 donors and use different weighting from Table 1. Bold denotes the largest supported-profile value in a row; the no-observation row is a candidate-prior diagnostic.

		Learned neural		Fixed
Observation support	Donors	R	$R+U$	Cosine $R+U$
Retained support	392	0.6224	**0.6454**	0.5204
Unresolved only	110	0.0000	0.1000	**0.1005**
No observations	55	0.0000	0.0727	0.0009

For donors with retained support, the neural full-context scorer exceeds both its retained-only counterpart and fixed cosine (Table 2). For unresolved-only donors, preserved context enables retrieval, but neural scoring offers no gain over cosine. With neither channel observed, every donor receives the same candidate ordering; nonzero hits reflect candidate priors rather than donor-specific recovery. This distinction prevents the value of retaining observations from being mistaken for a uniform benefit of learning.

The direction of the principal contrast survives stricter target confidence filtering. Requiring both synapse confidence scores to be at least 0.7 gives neural Hit@20 of 0.4642, versus 0.4306 for neural R and 0.4052 for fixed cosine. Absolute performance decreases as observations are removed. Appendix E reports the complete metric profile and the separate spatial-centroid references.

5 SUPPORTING EVIDENCE FOR THE VALUE OF UNRESOLVED RELATIONS

5.1 WITHHELD ADJACENCY PROVIDES A SEPARATE TEST OF INFORMATION

The revision task combines observation quality with the history of reconstruction edits. To test the information content of the unresolved channel through a different endpoint, we also evaluate masked adjacency. A symmetric unordered-pair mask removes both directions of a candidate relation before feature computation. On 128 prospectively selected MANC query neurons, 1,187,861 eligible pairs, and 12,675 observed positives, the learned retained-only scorer obtains macro AP 0.3800 and the matched segment-aware scorer obtains 0.4708 (Fig. 3). The strongest displayed fixed rule, forward resource allocation, reaches 0.4455. The observation contrast is therefore larger than the learned-over-fixed difference in this cohort.

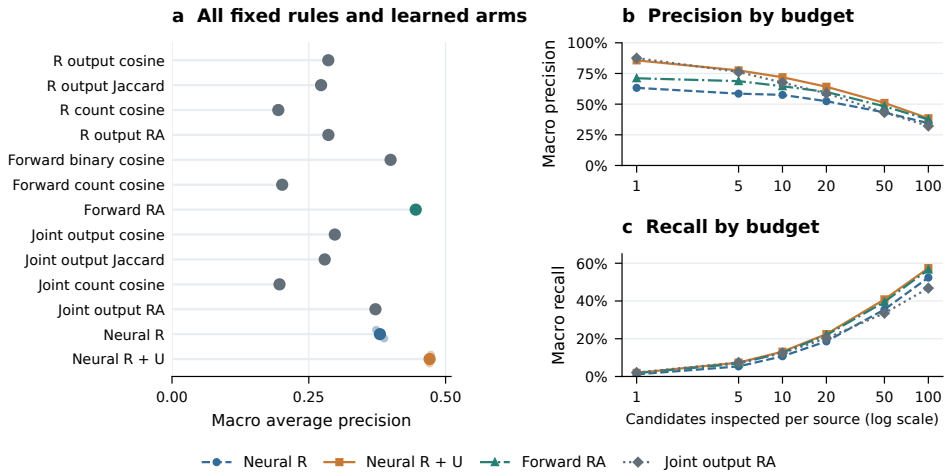

Figure 3: **Unresolved context improves full-list masked-adjacency retrieval. a,** Macro average precision for all eleven fixed rules and the two learned arms. **b–c,** Precision and recall versus inspection budget for selected comparators. All panels use the same 128-source cohort. Small dots in **a** show optimization seeds. RA denotes resource allocation.

A separate class-balanced control rewires binary incidence while preserving both degree sequences. Neural AP falls from 0.5206 to 0.4218–0.4228, with a similar decrease for logistic scoring. This result shows sensitivity to incidence structure beyond binary degree; weighted-count properties are not all preserved. The cohort differs from the full-list experiment, so the two AP ranges should not be compared as a single intervention trajectory. Appendix F records cohort sizes and the development history.

5.2 OBSERVATION VALUE AND MODEL CAPACITY HAVE DIFFERENT BOUNDARIES

Across the representation families studied, retaining unresolved observations improves AP, but the ranking of architectures depends on the input regime (Appendix F). Matched supervised ablations remove forward-path and common-output relation pools while preserving parameter shapes. Removing both pools reduces context-off AP to 0.030, whereas context-on AP remains 0.320. This contrast concerns access to relational evidence; the family comparisons also change training and architecture and cannot isolate that access effect as cleanly.

Structured loss provides a further boundary on learned combination. Training across five retained-output-loss levels improves intermediate-loss retrieval relative to original or endpoint-mixed training, but the gain varies by metric and loss level. At 75% loss on fresh target queries, multilevel neural training reaches AP 0.4318 versus 0.4146 for a shallow predictor trained on the same loss levels and 0.4152 for its intact-trained counterpart (Fig. 4); under intact input the shallow learner can be stronger. Predictive negative log likelihood is reported as a loss and does not, on its own, establish calibration. The matched shallow augmentation does not itself improve AP over intact training at 50% or 75%

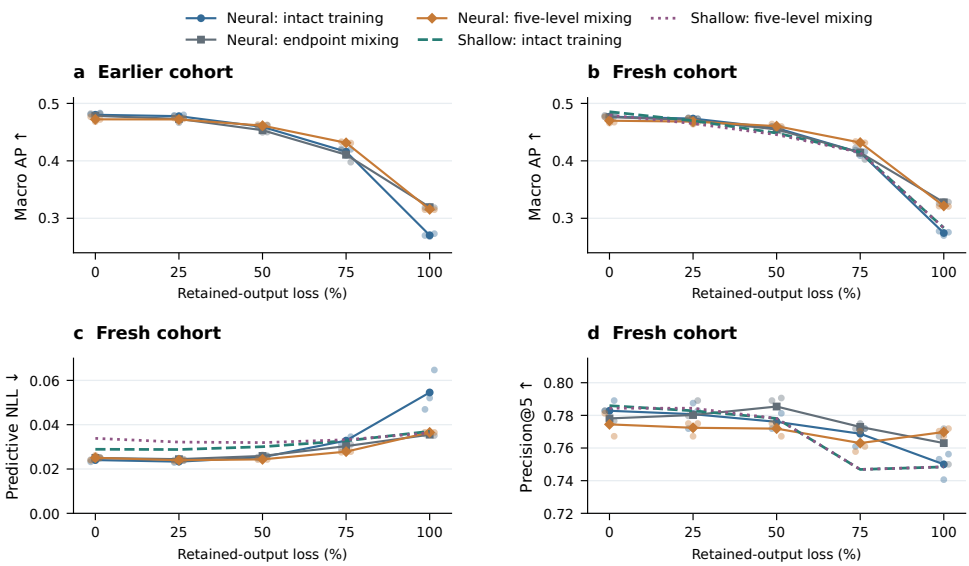

Figure 4: **The preferred learner changes with retained-output loss.** All arms retain segment context. **a,** Neural training regimes on an earlier query cohort. **b–d,** AP, predictive NLL, and Precision@5 on fresh queries, including both intact-trained and five-level-trained shallow comparators. Neural dots show three optimization seeds. Query cohorts share the MANC specimen; this is a synthetic-loss diagnostic.

loss. These comparisons control exposure to loss levels, while representation and optimization still differ. The cohort was used during method development; the result therefore characterizes a known synthetic-loss regime.

6 Related Work

Reconstruction and evolving identity. Connectomic learning increasingly operates on reconstructions that remain subject to correction. Geometry-based fragment linking uses point clouds (Berman et al., 2022) or multimodal volumetric features (Chen et al., 2024). NEURD focuses on morphology-aware decomposition and merge-error correction (Celii et al., 2025), whereas Autoproof learns from manual proofreading to support attachment and review workflows (Huang et al., 2025). ConnectomeBench evaluates proofreading decisions by vision-language models (Brown et al., 2025); ConnectomeBench2 extends this setting to large collections of expert-labeled split and merge decisions (Brown et al., 2026). Table 3 distinguishes their decision units from our full-list retrieval task. Their operational and multimodal results establish relevant capabilities, but different inputs and candidate-generation procedures prevent direct accuracy comparisons.

Relational prediction and representation. Probabilistic prediction from incomplete tract-tracing data already distinguishes observed evidence from unknown connectivity (Hinne et al., 2017). Connectomic representation learning also studies neuron-level circuit views (Li et al., 2024) and structure-to-activity prediction (Zhang et al., 2025); those endpoints differ from revision retrieval. In general link prediction, enclosing subgraphs (Zhang and Chen, 2018), structural sketches (Chamberlain et al., 2023), and learned paths (Zhu et al., 2021) capture complementary relational signals. Neural common-neighbor completion explicitly addresses incomplete structural context (Wang et al., 2024). Our comparison isolates access to a separately observed segment channel through a compact feature scorer; it does not benchmark all of these architectures.

Withholding the target during supervision. Denoising representation learning constructs a prediction target from corrupted input (Vincent et al., 2008). Leave-fragment-out supervision requires a particular form of withholding: sampled counts must be removed from the positive candidate

Table 3: **Related tasks and the decision they evaluate.** The comparison identifies scope, not a performance ranking. Citations and method details appear in the accompanying text.

Approach	Main evidence	Evaluated decision or output
Berman et al.	3D point clouds	Whether proposed fragments should connect
Chen et al.	Volume and morphology	Fragment connection for neuron tracing
NEURD	Meshes and synapses	Decomposition, merge correction, and features
Autoproof	Proofreading features	Prioritization and automatic attachment
ConnectomeBench2	EM images and meshes	Split correction and merge identification
This study	Directed partner profiles	Ranking every earlier traced candidate for a later recorded donor destination

before all dependent statistics are recomputed. This addresses the exact additive shortcut in Eq. (4), while leaving repeated-partner evidence available. Appendix H distinguishes this construction from blind denoising, contrastive learning, and other graph encoders. Our sampled-list objective follows the ranking perspective of Cao et al. (2007); it uses listwise normalization rather than the pairwise preference objective of Rendle et al. (2009).

7 Discussion and conclusion

Preserve evidence while identity remains provisional. The observation graph makes a useful distinction between the existence of a measured relation and the completeness of its endpoint identity. Across the evaluated tasks, unresolved incidence contains information that is discarded by a traced-neuron projection. The source-learning experiment shows that part of this information can improve retrieval in another connectome after the query has been removed from its synthetic owner. The contribution lies in the combination of representation, supervision construction, and evaluation: each is needed to interpret the measured gain.

Use the model at the operating point supported by the evidence. The improvement is a coverage result at $K = 20$. Lower Hit@1, Hit@5, and MRR mean that the scorer is less suitable when only the first few candidates can be inspected. The support analysis further separates two uses of context: unresolved-only observations enable retrieval with a fixed rule, while learned combination is more useful in the retained-support stratum. A practical retrieval stage should therefore specify its budget and available evidence before model selection, and retain geometry or image review for the final structural decision. The present results do not quantify the cost or acceptance rate of that workflow.

Limits of transfer and validation. The study uses two reconstructed individuals with different anatomical coverage and confidence definitions. Its revision cohort is conditioned on recorded many-to-one convergence and an earlier traced destination; it does not cover every split, merge, or missing candidate. Synthetic count fragments need not resemble contiguous neurites, and recorded destinations can inherit reconstruction errors. Although the final fits use source-only supervision, historical target feedback informed task development. This is consequently a transfer evaluation within an established research process, not a previously untouched external test. Optimization seeds and component-removal analyses characterize fit variation and finite-cohort breadth, without supplying biological population uncertainty.

Next validation and conclusion. A prospective test should freeze the release, candidates, model family, and review budget before inspecting subsequent outcomes, then measure accepted edits and expert time. Image and mesh comparisons should share the same candidate-generation and evaluation procedure. The present evidence supports preserving unresolved observations for relational inference and identifies an additional, budget-dependent benefit of residualized source supervision.

AI USE STATEMENT

AI tools were used for code development.

REPRODUCIBILITY STATEMENT

Sections 2–3 specify the representation, fragment law, residualization, features, objective, freezing, ties, and aggregation. The appendix gives derivations, cohort construction, comparisons, interventions, and implementation boundaries. The accompanying manuscript package includes editable figure sources and aggregate data for the redesigned result figures. Portable feature construction, frozen inference, and evaluation exist in the research materials; the manuscript package alone is not a complete raw-data-to-training release.

REFERENCES

Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 524–533. PMLR, 2019. URL https://proceedings.mlr.press/v97/batson19a.html.

Stuart Berg et al. Sexual dimorphism in the complete Drosophila male central nervous system connectome. *Cell*, 189:5504–5526.e15, 2026. doi: 10.1016/j.cell.2026.08.015.

Jules Berman, Dmitri B. Chklovskii, and Jingpeng Wu. Bridging the gap: Point clouds for merging neurons in connectomics. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pages 150–159. PMLR, 2022. URL https://proceedings.mlr.press/v172/berman22a.html.

Jeff Brown, Andrew Kirjner, Annika Vivekananthan, and Edward S. Boyden. ConnectomeBench: Can LLMs proofread the connectome? In *Advances in Neural Information Processing Systems 38*, volume 38, 2025. doi: 10.52202/085713-1728. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4a70f0c8443593ca59a88ebc8a937ed6-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Jeff Brown, Tim Farkas, Gleb Razgar, and Edward S. Boyden. ConnectomeBench2: A unified benchmark for automated connectomic proofreading. arXiv:2606.21116, 2026. URL https://arxiv.org/abs/2606.21116.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *International Conference on Machine Learning*, pages 129–136, 2007. doi: 10.1145/1273496.1273513.

Brendan Celii et al. NEURD offers automated proofreading and feature extraction for connectomics. *Nature*, 640:487–496, 2025. doi: 10.1038/s41586-025-08660-5.

Benjamin Paul Chamberlain, Sergey Shirobokov, Emanuele Rossi, Fabrizio Frasca, Thomas Markovich, Nils Hammerla, Michael M. Bronstein, and Max Hansmire. Graph neural networks for link prediction with subgraph sketching. In *International Conference on Learning Representations*, 2023.

Qihua Chen, Xuejin Chen, Chenxuan Wang, Yixiong Liu, Zhiwei Xiong, and Feng Wu. Learning multimodal volumetric features for large-scale neuron tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1174–1182, 2024. doi: 10.1609/aaai.v38i2.27879. URL https://ojs.aaai.org/index.php/AAAI/article/view/27879.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Sven Dorkenwald, Casey M. Schneider-Mizell, et al. CAVE: Connectome annotation versioning engine. *Nature Methods*, 22:1112–1120, 2025. doi: 10.1038/s41592-024-02426-z.

Sven Dorkenwald et al. Neuronal wiring diagram of an adult brain. *Nature*, 634:124–138, 2024. doi: 10.1038/s41586-024-07558-y.

Daniel Haehn, Verena Kaynig, James Tompkin, Jeff W. Lichtman, and Hanspeter Pfister. Guided proofreading of automatic segmentations for connectomics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9319–9328, 2018. doi: 10.1109/CVPR.2018.00971. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Haehn_Guided_Proofreading_of_CVPR_2018_paper.html.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, pages 1024–1034, 2017.

Max Hinne, Annet Meijers, Rembrandt Bakker, Paul H. E. Tiesinga, Morten Mørup, and Marcel A. J. van Gerven. The missing link: Predicting connectomes from noisy and partially observed tract tracing data. *PLOS Computational Biology*, 13(1):e1005374, 2017. doi: 10.1371/journal.pcbi.1005374.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-supervised masked graph autoencoders. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604, 2022. doi: 10.1145/3534678.3539321. URL https://dl.acm.org/doi/10.1145/3534678.3539321.

Gary B. Huang, William M. Katz, Stuart Berg, and Louis Scheffer. Autoproof: Automated segmentation proofreading for connectomics. arXiv:2509.26585, 2025. URL https://arxiv.org/abs/2509.26585.

Michal Januszewski et al. High-precision automated reconstruction of neurons with flood-filling networks. *Nature Methods*, 15:605–610, 2018. doi: 10.1038/s41592-018-0049-4.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Thomas N. Kipf and Max Welling. Variational graph auto-encoders. arXiv:1611.07308, 2016. URL https://arxiv.org/abs/1611.07308.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2Void - learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2129–2137, 2019. doi: 10.1109/CVPR.2019.00223. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Krull_Noise2Void_-_Learning_Denoising_From_Single_Noisy_Images_CVPR_2019_paper.html.

Junchi Li, Guojia Wan, Minghui Liao, Fei Liao, and Bo Du. Self-supervised contrastive graph views for learning neuron-level circuit network. In *Medical Image Computing and Computer Assisted Intervention*, pages 590–600, 2024. doi: 10.1007/978-3-031-72120-5_55.

Zudi Lin, Donglai Wei, Jeff Lichtman, and Hanspeter Pfister. PyTorch Connectomics: A scalable and flexible segmentation framework for EM connectomics. arXiv:2112.05754, 2021. URL https://arxiv.org/abs/2112.05754.

MICrONS Consortium. Functional connectomics spanning multiple areas of mouse visual cortex. *Nature*, 640(8058):435–447, 2025. doi: 10.1038/s41586-025-08790-w.

Stephen M. Plaza et al. neuprint: An open access tool for EM connectomics. *Frontiers in Neuroinformatics*, 16:896292, 2022. doi: 10.3389/fninf.2022.896292.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Conference on Uncertainty in Artificial Intelligence*, pages 452–461, 2009.

Louis K. Scheffer et al. A connectome and analysis of the adult Drosophila central brain. *eLife*, 9: e57443, 2020. doi: 10.7554/eLife.57443.

Philipp Schlegel et al. Whole-brain annotation and multi-connectome cell typing of Drosophila. *Nature*, 634:139–152, 2024. doi: 10.1038/s41586-024-07686-5.

Alexander Shapson-Coe et al. A petavoxel fragment of human cerebral cortex reconstructed at nanoscale resolution. *Science*, 384:eadk4858, 2024. doi: 10.1126/science.adk4858.

Shin-ya Takemura et al. A connectome of the male Drosophila ventral nerve cord. *eLife*, 13:RP97769, 2024. doi: 10.7554/eLife.97769.1. Reviewed Preprint, Version 1.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018. URL https://arxiv.org/abs/1807.03748.

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019.

Petar Veličković et al. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pages 1096–1103, 2008. doi: 10.1145/1390156.1390294.

Xiyuan Wang, Haotong Yang, and Muhan Zhang. Neural common neighbor with completion for link prediction. In *International Conference on Learning Representations*, 2024.

Michael Winding et al. The connectome of an insect brain. *Science*, 379:eadd9330, 2023. doi: 10.1126/science.add9330.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Zihan Zhang, Huanqiu Zhang, and Stefan Mihalas. Predicting neural activity from connectome embedding spaces. bioRxiv preprint, 2025. URL https://www.biorxiv.org/content/10.1101/2025.05.09.653224v1. Presented at the NeurIPS 2025 Workshop: Data on the Brain & Mind.

Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. Neural Bellman–Ford networks: A general graph neural network framework for link prediction. In *Advances in Neural Information Processing Systems*, 2021.

Jonathan Zung, Ignacio Tartavull, Kisuk Lee, and H. Sebastian Seung. An error detection and correction framework for connectomics. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/4500e4037738e13c0c18db508e18d483-Paper.pdf.

A FORMAL PROPERTIES OF LEAVE-FRAGMENT-OUT SUPERVISION

A.1 SELF-OVERLAP DECOMPOSITION AND BOUNDS

For a nonnegative integer owner profile x and a sampled fragment $0 \le q \le x$, write $r = x - q$. The unmodified positive-candidate dot product decomposes exactly as

$$q^\top x = q^\top r + \|q\|_2^2. \tag{8}$$

If $\sum_j q_j = m$, then $q_j^2 \ge q_j$ gives $\|q\|_2^2 \ge m$, while nonnegativity gives $\|q\|_2^2 \le (\sum_j q_j)^2 = m^2$. The shortcut is largest when sampled counts concentrate on few coordinates. Subtraction must precede every candidate-dependent feature: cosine changes its denominator, Jaccard changes when the final count at a coordinate is removed, and resource allocation changes because the residual candidate document frequencies change.

A.2 EXPECTED RESIDUAL SIGNAL UNDER FINITE-POPULATION SAMPLING

Let $n = \sum_j x_j > 1$ and sample q by Eq. (3), with $1 \le m \le n$. Multivariate-hypergeometric moments give

$$\mathbb{E}[q_j] = \frac{m x_j}{n}, \qquad \mathbb{E}[q_j(q_j - 1)] = \frac{m(m-1)x_j(x_j-1)}{n(n-1)}. \tag{9}$$

The second identity follows by counting ordered pairs of distinct population endpoints at coordinate j. Since $q_j^2 = q_j(q_j - 1) + q_j$,

$$\boxed{\mathbb{E}[q^\top(x-q)] = \frac{m(n-m)}{n(n-1)} \sum_j x_j(x_j - 1).} \tag{10}$$

By contrast, $\mathbb{E}[q^\top x] = (m/n)\sum_j x_j^2$. Residual same-coordinate evidence comes from repeated endpoint counts at shared directed partners. For a binary profile, $q^\top(x-q) = 0$ for every valid fragment. This is an identifiability limit of the synthetic count task, not evidence that real spatial fragments lack shared partners. These identities do not determine normalized-feature performance or establish neural superiority.

A.3 UNSUPPORTED DONORS REDUCE TO A CANDIDATE PRIOR

If donors d and d' have no observations in either channel, their relation and query-strength features are identical for every candidate; only candidate-strength coordinates can vary. Under frozen normalization and parameters,

$$q_d = q_{d'} = 0 \implies \phi(d,c) = \phi(d',c) \implies s_\theta(d,c) = s_\theta(d',c), \qquad \forall c. \tag{11}$$

Any nonzero hit therefore measures agreement between a shared candidate ordering and recorded destinations, not donor-specific connectivity recovery. If all scores tie, Eq. (6) reduces to $\mathrm{Hit}_K = \min(K, |\mathcal{C}|)/|\mathcal{C}|$.

A.4 TIE INTEGRATION

Let a candidates score strictly above the destination and b candidates, including the destination, share its score. Uniformly random ordering inside the tie block gives the probability that the destination occupies a top-K position,

$$\mathrm{Hit}_K = \frac{1}{b}\sum_{r=a+1}^{a+b} \mathbf{1}[r \le K] = \frac{\min\{b, \max(0, K-a)\}}{b}. \tag{12}$$

The reciprocal-rank expression in Eq. (6) is the same tie-block expectation. Deterministic ID-based tie breaking would make metrics depend on arbitrary body numbering.

B COMPLETE TASK CONSTRUCTION AND ALGORITHM

B.1 RECORDED-REVISION COHORT

The two MANC partner exports contain 86,952,391 rows each. Exact matching on six pre/post coordinates finds no duplicate keys and excludes six coordinate-inconsistent rows. The remaining 18,072 changed-assignment rows induce 41 many-to-one and 11 many-to-many components. We select the 35 many-to-one components whose same-ID destination was already traced in the old annotation. Their 557 donors comprise 495 IDs absent from the annotation table and 62 with null status; these are export metadata categories, not a certified biological fragment ontology.

All 23,200 old traced candidates and all donors remain in every comparison, including unsupported cases. The full partner space contains 18,360,982 columns per direction, including explicit zero columns required to preserve candidate membership. New-release assignments are used only as outcomes. The task does not evaluate image segmentation, expert acceptance, donor detection when the destination is outside the candidate set, or physiological function.

B.2 FEATURE MAP

For nonnegative vectors a, b, let $B(a) = \{j : a_j > 0\}$ and define cosine as zero if either norm is zero. For channel H and residual candidate collection $\{z_c^H\}_{c\in\mathcal{C}_s}$,

$$J(a,b) = \frac{|B(a)\cap B(b)|}{|B(a)\cup B(b)|}, \qquad A(a,b) = \sum_{j\in B(a)\cap B(b)} \frac{1}{f_j^H}, \qquad f_j^H = \sum_{c\in\mathcal{C}_s} \mathbf{1}[z_{c,j}^H > 0]. \tag{13}$$

Empty unions yield zero. Each channel contributes outgoing, incoming, and joint cosine; J; A; and log outgoing/incoming query and candidate strengths. Incoming and outgoing coordinates remain distinct even for the same partner ID.

B.3 TRAINING AND FROZEN TRANSFER

1. Split 256 source owners into 192 fitting and 64 validation owners; retain all 4,096 source candidates.
2. Draw four fixed fragments per owner, one at each $m \in \{1, 2, 4, 8\}$, and subtract each query from its positive candidate.
3. Build 96 fixed negatives per task from the top 32 retained and top 32 unresolved joint-cosine candidates, then deterministic uniform filling.
4. Fit 18-feature source normalization on fitting pairs only. The context-off arm zeros the nine standardized unresolved-channel coordinates.
5. Train linear and neural scorers for 200 Adam updates with 16 owner/fragment samples per update, learning rate 0.003, weight decay 10^{-4}, and seeds 17, 29, and 43.
6. Freeze all twelve final checkpoints and source normalization. Build old-release MANC profiles, score all 23,200 candidates for each donor, integrate ties, and aggregate first within component and then across components.

No validation-selected epoch or target-selected seed is used. Training with 96 negatives per query does not make Eq. (5) an unbiased full-candidate softmax.

C COMPLETE REVISION-RETRIEVAL COMPARISONS

Table 4: Fixed connectivity rules under retained-only versus all-partner profiles. Candidate membership and labels are unchanged. Bold marks the best rule in each column. H1 and H20 denote Hit@1 and Hit@20.

	Hit@1		Hit@20	
Rule	R	$R+U$	R	$R+U$
Outgoing cosine	0.0921	0.0607	0.3076	0.2872
Incoming cosine	0.1045	0.1323	0.3801	0.3974
Joint cosine	**0.1264**	**0.1415**	**0.4519**	**0.4666**
Joint Jaccard	0.0330	0.0626	0.4500	0.2769
Joint resource allocation	0.0481	0.0482	0.3013	0.2790
Strength prior	0.0000	0.0000	0.0000	0.0000

Table 5: Target-component-trained diagnostic, using five folds that each hold out seven components. Three-seed means. This setup does not beat fixed all-partner cosine at Hit@20 and changes both supervision domain and training-set size relative to source learning.

Diagnostic arm	Hit@1	Hit@5	Hit@20	MRR
Linear, R	**0.0981**	**0.2702**	**0.3496**	**0.1724**
Linear, $R+U$	0.0776	0.2554	0.3201	0.1542
Neural, R	0.0935	0.2306	0.3409	0.1591
Neural, $R+U$	0.0794	0.2153	0.3410	0.1469

D CORRESPONDENCE INTERVENTIONS, COVERAGE, AND SUPPORT

Table 6: Frozen source-trained context-on models under unresolved-relation interventions. Difference is observed minus intervened component-macro Hit@20, averaged across the same three seeds. Permutations preserve each query's multiset of five relation features; strength features do not move. Differences are computed before rounding.

Family / intervention	Observed	Intervened	Difference
Linear / shuffle 1	0.5035	0.4753	0.0283
Linear / shuffle 2	0.5035	0.4726	0.0310
Linear / shuffle 3	0.5035	0.4726	0.0310
Linear / zero relations	0.5035	0.4763	0.0272
Neural / shuffle 1	0.5399	0.4930	0.0469
Neural / shuffle 2	0.5399	0.4907	0.0492
Neural / shuffle 3	0.5399	0.4933	0.0466
Neural / zero relations	0.5399	0.4964	0.0435

Table 7: Component coverage of seed-mean neural $R + U$ Hit@20 differences over 35 components. "Remove top k" discards the k largest positive component differences before averaging the rest.

			Gain after removal		
Comparator	Full gain	Better/worse/tie	Top 1	Top 3	Top 5
Neural R	0.0533	18/4/13	0.0450	0.0309	0.0172
Linear $R + U$	0.0363	17/2/16	0.0276	0.0182	0.0120
Fixed cosine	0.0733	19/11/5	0.0637	0.0494	0.0387

Table 8: Donor-average Hit@20 by available observation support. The weighting differs from component-macro results. The three strata partition all 557 donors.

Support	Donors	Neural R	Neural $R + U$	Linear $R + U$	Fixed cosine
No observations	55	0.0000	0.0727	0.0000	0.0009
Outside only	110	0.0000	0.1000	0.0848	**0.1005**
Retained support	392	0.6224	**0.6454**	0.6131	0.5204

The no-observation row is a prior diagnostic, so its maximum is not highlighted. The no-profile neural hits occur only for seed 17 and come from twelve donors in one component. This is consistent with a shared candidate prior and does not contradict the proposition above.

E CONFIDENCE AND SPATIAL SENSITIVITY

Table 9: Component-macro retrieval when both target confidence scores are at least 0.7. Models, labels, candidates, and source normalization are frozen. Bold marks the maximum in each metric.

Scorer	Hit@1	Hit@5	Hit@20	MRR
Linear R	0.1083	0.2787	0.4330	0.1869
Linear $R + U$	0.0825	0.2579	0.4411	0.1597
Neural R	0.1104	**0.2915**	0.4306	**0.1928**
Neural $R + U$	0.0787	0.2611	**0.4642**	0.1593
Fixed cosine	**0.1108**	0.2681	0.4052	0.1819

Table 10: Synapse-centroid references on the same revision cohort. Coordinates are partner-row-weighted summaries, not soma positions, arbor extent, mesh adjacency, or image evidence.

Centroid	Hit@1	Hit@5	Hit@20	MRR
Outgoing	0.0056	0.0119	0.0257	0.0132
Incoming	**0.0170**	**0.0259**	**0.0562**	**0.0290**
Pooled	0.0125	0.0224	0.0527	0.0265

F SUPPORTING OBSERVATION-LEARNING COHORTS AND FIGURES

The main paper deliberately separates cohorts that answer different questions. The full-list masked-adjacency cohort has 128 sources, 1,187,861 candidate pairs, and 12,675 observed positives. The representation comparison has 1,187,608 pairs. The class-balanced degree-preserving control has 1,186,818 pairs and 9,280 positives. The fresh structured-loss cohort has 128 query sources, 1,189,146

pairs, and 10,505 positives. Both intact-trained and five-level-trained shallow controls use this cohort alongside the neural models in Fig. 4. The historical restricted-budget cohort below has 383,076 pairs. These are not interchangeable replications.

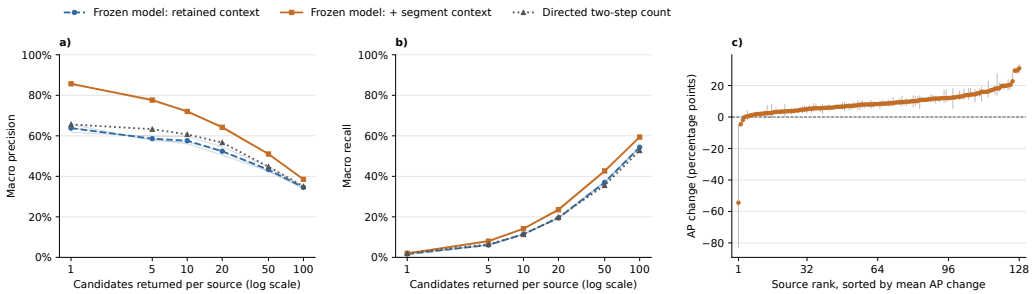

Figure 5: Restricted-candidate masked-adjacency analysis on a distinct 128-source cohort. Precision, recall, and source-wise AP changes are shown. At budget 100 one source has only 50 candidates. This is not the 1.19-million-pair full-list experiment.

The development sequence in Fig. 6 records within-setting comparisons across successive analyses. Its rows have different candidate universes and cannot be pooled into a single effect estimate. Figure 7 separates family comparisons from matched supervised pool ablations.

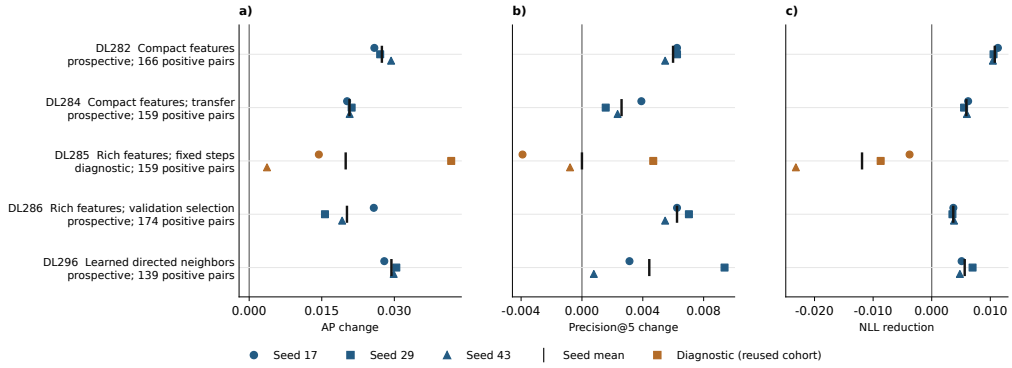

Figure 6: **Development history of paired context effects.** Blue markers are prospective seeds and black bars are means; orange markers indicate diagnostic cohort reuse. Panels report AP change, Precision@5 change, and NLL reduction. Comparisons are within rows; cohort reuse and different candidate universes prevent treating the rows as independent replications.

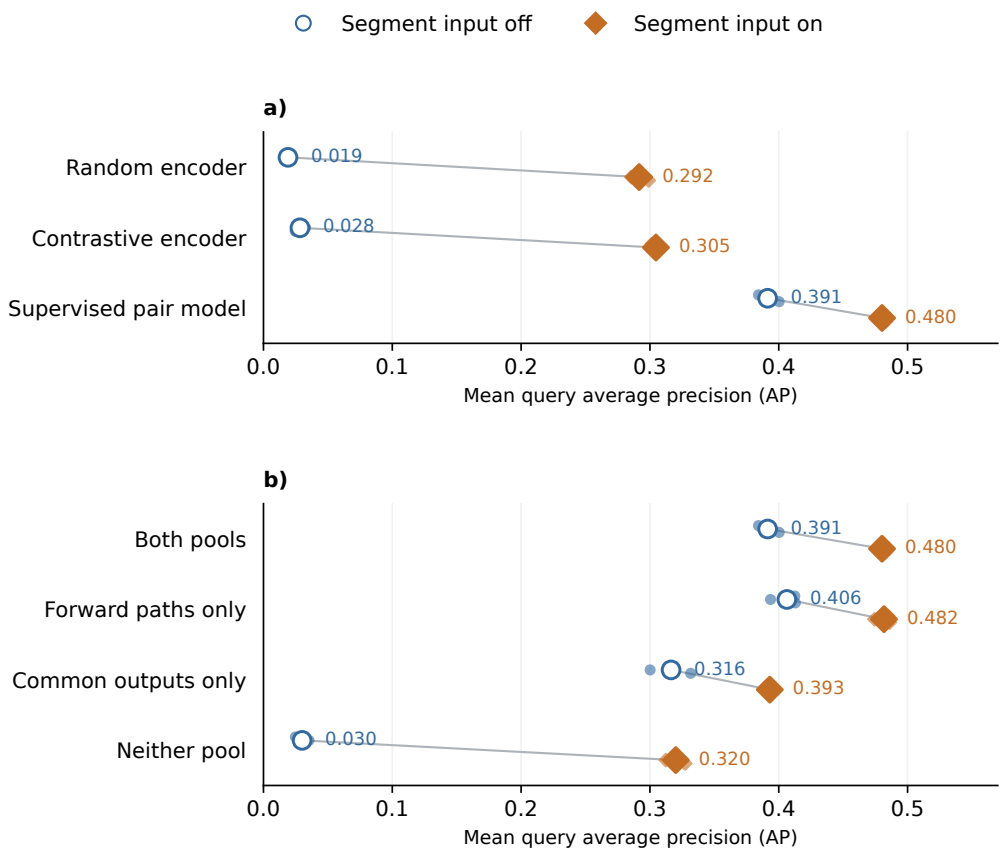

Figure 7: **Representation comparisons and matched pool ablations.** Top: families differ in architecture and training. Bottom: supervised relation-pool ablations preserve parameter shapes and retrain after input removal. Context-off and context-on arms differ in access to unresolved observations.

G Implementation and interpretation boundaries

The resource-allocation denominator is recomputed from residual candidates for every synthetic task. The four fragment sizes are independent hypothetical tasks, not four disjoint pieces of one reconstructed neuron. Target candidates remain present even when their profiles are zero. Featurization over the candidate universe dominates end-to-end cost; the 1,153-parameter scorer is cheap only after those statistics exist.

Seeds, queries, donor segments, and revision components do not increase the number of biological specimens. Seed variation is not a biological confidence interval. Component-removal diagnostics establish finite-cohort breadth, not population significance. A component-level bootstrap or a new budget-integrated statistic would require the frozen per-component predictions and a predeclared analysis; it cannot be reconstructed faithfully from rounded manuscript means.

The two connectomes have different anatomical coverage and native confidence definitions. Historical target feedback informed task development even though the final fits are source-only. Any subsequent external test should freeze the release, candidate set, budget, and model family before target outcome inspection. Comparisons with image, mesh, or multimodal systems require matched candidate generation and evaluation rather than juxtaposition of unrelated headline accuracies.

H BROADER METHODOLOGICAL CONTEXT

From reconstruction to queryable observations. Adult central-brain (Scheffer et al., 2020) and larval-brain reconstructions (Winding et al., 2023) illustrate the diversity of available insect connectomes. Mouse visual-cortex reconstructions link structural and functional measurements (MICrONS Consortium, 2025), while nanoscale human-cortex data extend the scale of mammalian reconstruction (Shapson-Coe et al., 2024). Across these settings, annotation and cross-connectome cell typing address identity at a different level from local segment retrieval (Schlegel et al., 2024). Tools such as neuPrint support structured access to connectivity and annotations (Plaza et al., 2022). Our retained/unresolved partition is a selection within observed data, not a claim that a particular release exhausts the biological circuit.

Reconstruction tools and correction workflows. Flood-filling networks learn volumetric segmentation (Januszewski et al., 2018); PyTorch Connectomics provides a framework for EM segmentation models (Lin et al., 2021). Error detection and correction (Zung et al., 2017) and guided proofreading (Haehn et al., 2018) evaluate intervention points in reconstruction workflows. These tasks can supply candidates or complementary evidence, but their reported metrics do not directly measure our all-candidate destination-retrieval endpoint.

Graph encoders. Neighborhood sampling in GraphSAGE (Hamilton et al., 2017), graph convolution (Kipf and Welling, 2017), attention (Veličković et al., 2018), and injective neighborhood aggregation (Xu et al., 2019) offer different choices of encoder. Variational graph autoencoding reconstructs relational structure (Kipf and Welling, 2016), while GraphMAE learns from masked graph attributes (Hou et al., 2022). Deep Graph Infomax uses agreement between local representations and a graph summary (Veličković et al., 2019). These precedents motivate representation learning on graphs; they do not establish that a deeper encoder would improve the specific observation contrasts studied here.

Self-supervision and withholding. Noise2Self (Batson and Royer, 2019) and Noise2Void (Krull et al., 2019) use withheld observations for blind denoising under their respective assumptions. SimCLR constructs paired augmented views for contrastive learning (Chen et al., 2020), and contrastive predictive coding learns representations through prediction of latent observations (van den Oord et al., 2018). The count-fragment construction shares the broad practice of obtaining supervision from available data. Its guarantee is narrower: subtraction removes the exact self-overlap term in Eq. (4). It does not inherit noise-independence guarantees or establish equivalence to spatial fragment recovery.