

Learning the proofreading decision: data scaling, transfer to expansion microscopy, and a pre-registered test of automatic repair on nine held-out regions of a mouse cortex connectome

Dan Siroker¹

¹Big Brain Lab. Correspondence: bigbrainlab.ai

Abstract. Connectome reconstruction is limited by human proofreading: the automated tracing of electron-microscopy volumes leaves breaks and wrong joins that people repair by hand, at a cost projected to dominate the budget of a whole mouse brain⁷. We ask four questions. Can the proofreading decision be learned from logged corrections, and how does its accuracy scale with data? Does what is learned on electron microscopy transfer to expanded light-microscopy tissue? Does a pipeline built from such models repair a released segmentation when tested once, on regions it has never seen? And can candidate repairs be proposed without proofread labels at all? A two-stream convolutional network of 6.9 million parameters trained on ConnectomeBench2 decisions reaches 97.4% balanced accuracy on split correction and 94.4% on merge identification (three seeds; 101,929 held-out decisions), above the benchmark’s published expert point estimates (93.0, 84.1) and level with or above its published Vision Transformer (97.0, 93.0), which trained on the benchmark’s training split from ImageNet weights. Over the 1.9 decades of the sweep, error falls as a power law in training decisions with exponents of 0.20 and 0.16, and runs on all 531,734 downloaded decisions landed 0.4 and 2.7 percentage points above the extrapolation. Without retraining, the same model scores AUROC 0.70 and 0.76 on E11 Bio’s barcoded expansion-microscopy data, and at 250 barcode labels fine-tuning from it matches a generic initialization given four to eight times the labels. In a pre-registered test on nine never-seen regions of the MICrONS minnie65 volume (1,754 proofread neurons), the primary comparison failed: an image-based second stage changed normalized expected run length by +0.003 [-0.005, +0.012], below its pre-specified criterion. In a secondary comparison, a pipeline of a pair model and a tip judge raised run length from 0.818 to 0.858 (+0.040 [+0.027, +0.054]), positive in every region by point estimate, with breaks down 22% and wrong joins up 3%. Because 95.4% of applied joins connect tissue neither side of which has been proofread, their precision is not measured; designing that measurement is the open problem the results leave.

Introduction

Synaptic-resolution reconstructions now include the adult fruit fly brain^{1,2}, a cubic millimeter of mouse visual cortex³, a cubic millimeter of human temporal cortex⁴ and, reported in 2026, the complete male fly central nervous system⁵. Each combined automated segmentation, most recently by flood-filling networks⁶, with human proofreading of the breaks (false splits) and wrong joins (false merges) the segmentation leaves behind. Proofreading is the cost that scales. The fly connectome took a community of hundreds of proofreaders several years; the cubic millimeter of mouse cortex was a multi-year, multi-institution program whose reconstruction remains partly proofread; and a 2023 Wellcome projection, under that report’s assumptions, put a whole mouse brain at \$7.5 to 21.7 billion all-in, most of it proofreading⁷.

Two developments make the question of automating the decision timely. First, the corrections themselves are now logged. The Connectome Annotation Versioning Engine⁸ records every proofreader edit in the fly, mouse, human and zebrafish reconstructions, and ConnectomeBench2⁹ packages 716,485 of those decisions, with negative controls, as a supervised benchmark with published human-expert figures on the same task format. Second, the tissue is changing. Light-microscopy connectomics of expanded tissue¹⁰ and whole expanded mouse brains with sparsely labeled neurons imaged in as little as a day¹¹ promise maps at a fraction of the cost of electron microscopy, and protein barcodes, a separate method¹², offer a way to label cells so that a wrong join announces itself. Whether models trained on the electron-microscopy corrections of the

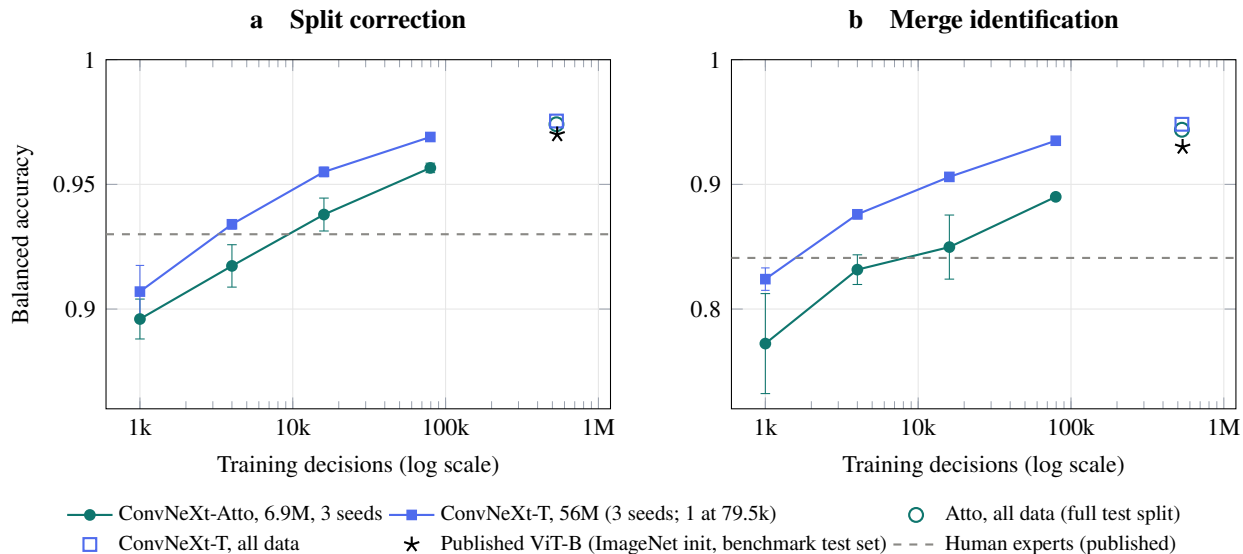


Figure 1: Accuracy of the learned proofreading decision against training data. Balanced accuracy on ConnectomeBench2’s held-out decisions. Filled markers: the 1,000- to 79,500-decision sweep, scored on 23,730 held-out decisions (bars are the seed standard deviation over three seeds; ConvNeXt-T at its tuned learning rate, three seeds at the first three points and one at 79,500). Open markers: the same architectures trained on all 531,734 downloaded decisions, scored on the full 101,929-decision test split (three seeds each). Star: the benchmark’s published Vision Transformer, trained from ImageNet weights on the benchmark’s 75% training split (about 537,000 of its 716,485 decisions) and scored on its held-out test set⁹, whose items may differ from ours. Dashed line: the benchmark’s published human-expert point estimate (93.0% and 84.1%; 95% intervals [87.2, 97.7] and [75.8, 91.7]). Power-law fits to the four Atto sweep points (1.9 decades) give error exponents of 0.201 (s.e. 0.010) and 0.158 (s.e. 0.017).

last decade carry over to that tissue determines how much of the proofreading problem starts solved.

Google’s PATHFINDER pipeline¹³ has reported an order-of-magnitude reduction in axon reconstruction error from automated agents that propose and check repairs. Here we ask a narrower set of questions with the public data and modest compute, and we answer each with a measurement we have tried to make hard to fool: a scaling sweep with three seeds per point (one at the largest ConvNeXt-T point) on a fixed held-out test set; a zero-shot and fine-tuning transfer test on barcoded expansion data with labels derived independently of the model; a pre-registered end-to-end test on nine regions sealed before the pipeline was frozen, scored once against the dataset’s own proofreading; and a label-free candidate stage compared against the same network trained with real labels. We report the pre-registered primary comparison, which failed, alongside the secondary comparisons, one of which (a label-free cleaner) also failed its pre-specified rule.

Results

The proofreading decision can be learned, and its accuracy follows a power law in data

ConnectomeBench2 poses two binary decisions: *split correction* (should these two segments be joined?) and *merge identification* (is this one segment wrongly joined from two cells?), each rendered as three orthographic views of the segment meshes and three orthogonal electron-microscopy slices at 224 pixels (Methods). We trained a two-stream ConvNeXt¹⁴, one backbone over the mesh views and one over the image slices, with one output per task, on stratified samples of 1,000 to 79,500 training decisions and on the full 531,734 we downloaded, and scored every run by balanced accuracy on held-out test decisions: a common 23,730-decision subset for the sweep and the full 101,929-decision test split for the full-data runs (Fig. 1, Table 1).

Table 1: Balanced accuracy against training decisions. Mean over three seeds (seed s.d. in parentheses) on the 23,730 held-out decisions of the sweep. ConvNeXt-T rows are at the tuned learning rate (5×10^{-5}), three seeds at the first three points and one at 79,500. Human and published-model figures are from the ConnectomeBench2 preprint⁹.

Model, task	Training decisions				
	1,000	4,000	16,000	79,500	531,734 [†]
ConvNeXt-Atto, split correction	89.6 (0.8)	91.7 (0.8)	93.8 (0.7)	95.7 (0.2)	97.4 (0.04)
ConvNeXt-Atto, merge identification	77.2 (4.0)	83.2 (1.2)	85.0 (2.6)	89.0 (0.04)	94.4 (0.06)
ConvNeXt-T, split correction	90.7 (1.0)	93.4 (0.2)	95.5 (0.2)	96.9	97.5 (0.10)
ConvNeXt-T, merge identification	82.4 (0.9)	87.6 (0.4)	90.6 (0.2)	93.5	94.8 (0.12)
Human experts, split correction (published)			93.0 [87.2, 97.7]		
Human experts, merge identification (published)			84.1 [75.8, 91.7]		
Published ViT-B (training split, about 537,000; ImageNet init)			97.0 / 93.0		

[†] Scored on the full 101,929-decision test split; the ConvNeXt-T full-data runs used learning rate 2×10^{-4} .

Three results follow. First, accuracy rises monotonically with data on both tasks and both model sizes. For the 6.9-million-parameter model the error, one minus balanced accuracy, is well described by a power law in training decisions over the 1.9 decades of the sweep from 1,000 to 79,500, with exponents of 0.201 (least-squares s.e. 0.010) on split correction and 0.158 (s.e. 0.017) on merge identification, fitted to the three-seed means. Each doubling of the data removes about 13% and 10% of the remaining error. The curve had not leveled off at the last point: the full-data runs, which trained for ten epochs rather than the sweep’s clamped schedule and were scored on the larger test split, landed 0.4 and 2.7 percentage points above the sweep’s extrapolation to 531,734 decisions. The 56-million-parameter ConvNeXt-T scales faster (exponents 0.253 and 0.225 at its tuned learning rate) and is ahead at every size from 1,000 decisions.

Second, at full data the models are at or above the published expert level on the same task format. ConvNeXt-Atto scores 97.41% (seed s.d. 0.04) on split correction, inside the human interval [87.2, 97.7], and 94.38% (s.d. 0.06) on merge identification, above the human interval [75.8, 91.7]. ConvNeXt-T scores 97.54 (s.d. 0.10) and 94.82 (s.d. 0.12). A DINOv3-initialized¹⁵ ConvNeXt-T at its own tuned learning rate reached 97.75 (s.d. 0.06) and 95.52 (s.d. 0.05) over three seeds, the best model we trained. “Above human” here means above the benchmark’s published expert point estimate, whose sample size the preprint does not state and whose intervals are wide, not a head-to-head on the same items.

Third, the published Vision Transformer is matched or beaten by much smaller models on about the same data. The benchmark’s ViT-B, a single encoder over both inputs trained from ImageNet weights on the benchmark’s training split, reports 97.0 and 93.0%. A two-stream ViT-B/16 of our own (two separate encoders, 172 million parameters) on our 531,734 decisions scored 97.29 and 93.90, so the published figure is within a point of what that architecture class gives on our pipeline. The three-seed Atto and ConvNeXt-T means exceed both: Atto by 0.1 and 0.5 percentage points over our ViT-B with 6.9 million parameters in total against 86 million in one ViT-B/16 encoder, at 0.38 times the wall clock (one Atto full-data run at a different learning rate, 97.22, landed below it on split correction). Across the twelve backbone and learning-rate settings at 79,500 decisions in Extended Data Table 1, the mid-sized models at a workable learning rate landed between 96.3 and 96.9 on split correction and 91.7 and 93.8 on merge identification; the learning rate mattered more than the backbone family.

The full-data Atto model is well calibrated (adaptive expected calibration error 0.008 and 0.014; AUROC 0.996 and 0.987) and scores on held-out decisions from each of the four species it was trained on: 98.7 and 96.0 on mouse, 95.8 and 87.5 on human, 97.0 and 90.5 on fly, 95.3 and 84.7 on zebrafish (split correction and merge identification; the human and zebrafish merge-identification test sets have 60 and 54 rows). For comparison, the best large language model on the earlier ConnectomeBench¹⁶ scored 75 to 79% accuracy on binary split correction and 62 to 63% on merge identification on that benchmark’s own items; the 1,000-

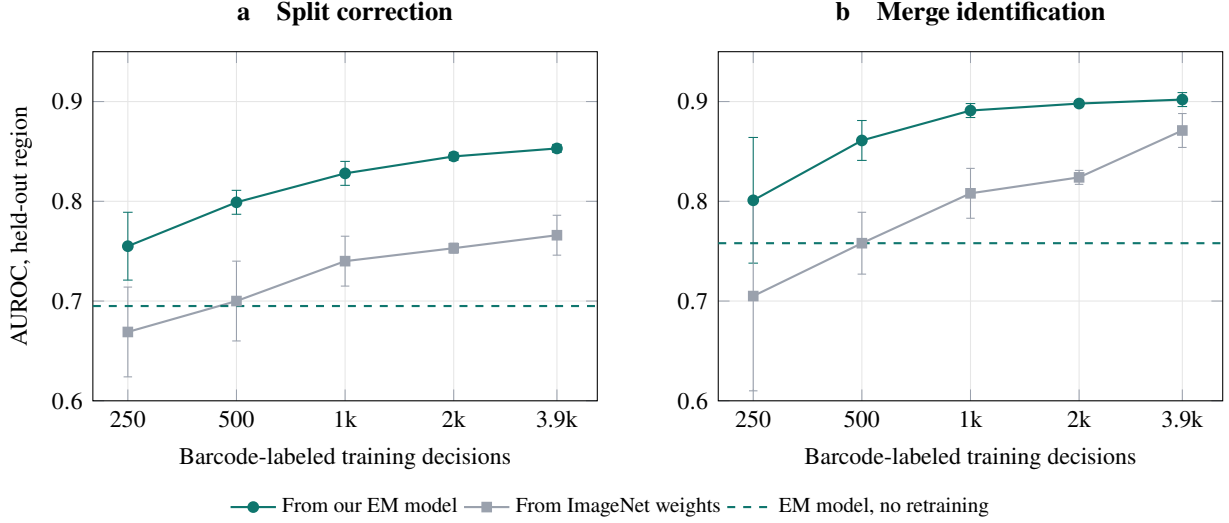


Figure 2: Transfer to barcoded expansion microscopy. AUROC on a held-out region of E11 Bio’s PRISM evaluation volume (564 split-correction and 672 merge-identification decisions, labels derived from protein barcodes the model never sees) against the number of barcode-labeled decisions used for fine-tuning; mean and s.d. over three seeds (30 runs). Green: fine-tuned from our electron-microscopy checkpoint. Gray: the same architecture from generic ImageNet weights. Dashed: the electron-microscopy checkpoint applied with no retraining (chance is 0.5). The power-law exponents quoted in the text are fits to balanced-accuracy error, not to the plotted AUROC.

decision Atto model (three-seed means 89.6 and 77.2 balanced accuracy, on different items) clears those figures by about 11 and 14 percentage points.

What is learned on electron microscopy transfers to barcoded expansion microscopy

E11 Bio’s PRISM data set is expansion microscopy of mouse hippocampus (CA3) in which each neuron carries a combinatorial protein barcode read out in 18 channels¹². The barcodes supply labels that no model sees: two touching segments with the same readable barcode are labeled the same cell, two with clearly different barcodes different cells, and a segment whose parts carry two barcodes a merge, under rules checked against human skeletons (Methods). We built the same two decision tasks on E11’s structure-only (GFP) segmentation of its evaluation region, rendered them in the ConnectomeBench2 format from the GFP channel and the segment meshes, with no barcode channel in any input, and held out a test region 31 μm (native) beyond the training region: 564 split-correction and 672 merge-identification decisions.

Applied with no retraining, the electron-microscopy Atto model scores AUROC 0.695 [0.652, 0.739] on split correction and 0.758 [0.713, 0.796] on merge identification, against 0.5 for chance and 0.99 on its own domain (Fig. 2, Table 2). Fine-tuned on 250 to 3,865 barcode-labeled decisions, three seeds per point, it improves monotonically on both tasks (0.755 to 0.853 and 0.801 to 0.902). Against the same architecture started from ImageNet weights it stays about 0.09 AUROC ahead on split correction at every size, while on merge identification the gap narrows from 0.10 at 250 labels to 0.03 at 3,865. Read off the three-seed means at the 250-label point, the electron-microscopy start matches the generic start with about 2,000 labels on split correction (0.755 against 0.753) and about 1,000 on merge identification (0.801 against 0.808): at that point the checkpoint is worth four to eight times the labels (a one-seed version of the same comparison had given 15 times, a seed-0 accident). Seed spread shrinks with data (balanced-accuracy s.d. 2.8 percentage points at 250 labels and 0.4 at 3,865 on split correction; 8.7 and 1.4 on merge identification), so the top of the curve is well determined. Power-law exponents on the error are positive with bootstrap intervals excluding zero on all four balanced-accuracy fits (0.131 [0.068, 0.198] and 0.222 [0.140, 0.312] from the electron-microscopy

Table 2: AUROC on E11 Bio’s PRISM held-out region. Mean over three seeds (seed s.d.). The zero-shot row is the electron-microscopy checkpoint with no retraining; chance is 0.5.

Barcode-labeled decisions	Split correction (564)		Merge identification (672)	
	From EM model	From ImageNet	From EM model	From ImageNet
0 (no retraining)	0.695	not run	0.758	not run
250	0.755 (0.034)	0.669 (0.045)	0.801 (0.063)	0.705 (0.095)
500	0.799 (0.012)	0.700 (0.040)	0.861 (0.020)	0.758 (0.031)
1,000	0.828 (0.012)	0.740 (0.025)	0.891 (0.007)	0.808 (0.025)
2,000	0.845 (0.004)	0.753 (0.005)	0.898 (0.003)	0.824 (0.007)
3,865 (all)	0.853 (0.004)	0.766 (0.020)	0.902 (0.007)	0.871 (0.017)

start). The released data cap the sweep at 1.2 decades, so whether the curve continues is not resolvable from them.

A check on cells the barcodes cannot label (seed 0 runs): 139 touching pairs in the test region are the same cell by human skeletons but have no usable barcode label (unreadable, a collision code, or a disagreement). Recall on them rose from 0.59 with no retraining to 0.68 at 250 labels and 0.85 at all 3,865 (electron-microscopy start), while specificity on the barcode test set was 0.79 at both 250 and 3,865 labels, so the gain between those two points did not come from a shift toward “same”; the zero-shot specificity was not recorded. The barcode labels cover about 13% of candidate pairs, biased toward bright, readable, non-colliding cells; this is the only check on the rest, and it covers positives only.

A pre-registered test of automatic repair on nine held-out regions

Scoring decisions on a benchmark is not repairing a volume. We therefore built a repair pipeline for the public MICrONS minnie65 segmentation³ and tested it once, under a protocol written down and hashed before the test regions were opened (Methods). The pipeline has three stages. A connected-components baseline relabels pieces that do not touch inside each padded evaluation crop; this corrects the start state for a cropped evaluation and is not counted as repair. A *pair model*, a three-dimensional network pre-trained on cuts made in the segmentation’s own segments (no proofread labels) and fine-tuned on 1,294 positive join examples from the proofread state of four other regions, scores candidate joins between segments, and a *tip judge* (“does this tip continue?”) retrained from the same four regions’ tip tables adds its rows; a fixed budget of 1,000 pair-model rows per region plus the judge’s rows defines state J3. The candidate under test, J4, adds image-based joins from a dense affinity network on 16×16×40 nm voxels. The pre-registered primary comparison was J4 against J3, with the success rule fixed in advance: a gain of at least 0.005 in plain normalized expected run length (nERL) and a paired 95% bootstrap lower bound above zero, with a larger-effect bar of +0.02 reported separately. An earlier draft had used +0.02 as the criterion; it was amended to +0.005 when the protocol was written, before any region was scored.

The nine evaluation regions were drawn by a written rule from a ranking of candidate regions by logged-edit density (densely proofread tissue, not a random sample of the volume), had never been used for training, tuning or selection, and lie at least 30 μ m from every region whose labels or cuts trained a model used here; two development regions on which pipeline choices were examined lie 20 μ m from regions B and E. They contain 1,754 proofread neurons (4,433 neuron-region records) with 345.5 mm of reference path on the preprint-rule skeletons used for run length and 388.9 mm on the standing skeletons used for break and wrong-join counts (Methods). Four further regions were sealed as a confirmation set and remain unopened. The pipeline’s inputs were read from a label-free cache; the proofreading truth, downloaded and skeletonized before the protocol was frozen, was read at test time only by the evaluator.

The primary failed. The image joins (232 across the nine regions, 9 to 47 each, against 20,578 J3 rows,

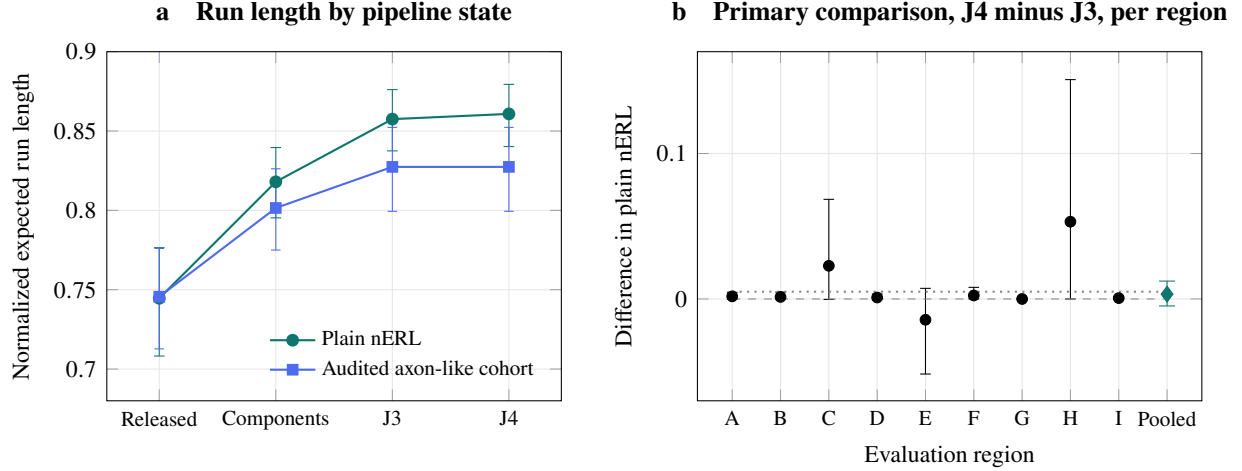


Figure 3: The pre-registered end-to-end test on nine held-out regions of MICrONS minnie65. **a**, Normalized expected run length of 1,754 proofread neurons under the released segmentation, the connected-components baseline, the pair-model-plus-tip-judge pipeline (J3) and the same with image-based joins (J4); bars are 95% bootstrap intervals over neurons grouped across regions (unpaired for the states in **a**, paired for the contrasts in **b**). Green: plain nERL on the preprint-rule reference skeletons (345.5 mm). Blue: the audited axon-like cohort on the same skeletons (82.8 mm classified axon-like), with wrong joins detected against every proofread skeleton. **b**, The pre-registered primary comparison, J4 minus J3, per region and pooled (diamond), with the pre-specified success threshold of +0.005 (dotted). The primary failed: +0.0033 [-0.0048, +0.0123]; the gain sits on one region and ten neurons.

so the stage touched about 1% of the join list) raised plain nERL by +0.0033 [-0.0048, +0.0123], below the 0.005 threshold and with an interval covering zero (Fig. 3, Table 3). The gain is concentrated: it sits on one region (H, +0.053) and on ten of 1,754 neurons; one neuron carries 59% of the positive contribution, four neurons lose, and without region H the pooled difference is -0.0002. On the audited axon-like cohort the image joins changed nothing in any region; on the all-cells audited evaluator the change was -0.0015 [-0.0081, +0.0026].

The secondary finding is improved run length on annotated processes. J3’s joins raised plain nERL from 0.818 to 0.858, +0.040 [+0.027, +0.054] over the components baseline, positive in every one of the nine regions by point estimate (per-region +0.011 to +0.109; region D’s interval covers zero; leave-one-region-out range +0.035 to +0.042), and raised the audited axon-like cohort by +0.026 [+0.017, +0.035]. Breaks on the 388.9 mm of standing-skeleton path fell from 2,826 to 2,198 (7.27 to 5.65 per mm; -22%), fewer in every region; wrong joins rose from 290 to 299 (0.75 to 0.77 per mm; +3%). From the released segmentation to J3 the plain run length rose by 0.113, of which the components convention accounts for 0.073.

Quality and scope. Of the 20,578 joins J3 applied across the nine regions, 269 can be decided against the finished-cell site truth: 238 correct and 31 wrong, a precision of 0.885 (Wilson 95% interval [0.841, 0.918]; pair-model rows 0.886, tip-judge rows 0.884). A further 148 rows are ambiguous (125 join different roots where neither is a finished cell, 23 have a side outside the site window); counting them as wrong gives 0.571. 410 rows join the same proofread but unfinished cell, 113 connect pieces of the same unproofread root (0.5%), and 19,638 rows, 95.4%, connect objects neither of which anyone has proofread (Table 4). The evaluation therefore measures run length on proofread cells, which is what the truth supports; precision on the rest is not measured. Volume belonging to no proofread cell but attached to proofread cells rose from 40,820 to 43,179 μm^3 under J3 (+5.8%), a proxy that counts legitimate unannotated continuations as well as errors.

A label-free cleaner (a supervoxel-shape classifier applied before J3) raised strict nERL (+0.017 [+0.011, +0.024]) and lowered the proxy volume by 8% but lowered the audited axon-like cohort (-0.008 [-0.016,

Table 3: Run length on the nine held-out regions, and the paired contrasts. All rows on the preprint-rule reference skeletons (345.5 mm). State rows carry 95% bootstrap intervals over neurons grouped across regions; contrast rows are paired bootstraps. The last column is the pre-registered all-cells audited evaluator, reported for completeness. Only J4 versus J3 was pre-registered as a comparison; the others are secondary analyses of the same frozen evaluation.

State	Plain nERL	Strict nERL	Audited axon-like	All cells, audited
Released segmentation	0.745 [0.708, 0.776]	0.224 [0.201, 0.248]	0.746 [0.713, 0.776]	0.769 [0.743, 0.793]
Components baseline	0.818 [0.795, 0.840]	0.384 [0.353, 0.415]	0.802 [0.775, 0.826]	0.826 [0.806, 0.844]
J3 (pair model + tip judge)	0.858 [0.837, 0.876]	0.389 [0.358, 0.422]	0.827 [0.799, 0.852]	0.860 [0.842, 0.877]
J4 (J3 + image joins)	0.861 [0.840, 0.879]	0.389 [0.358, 0.421]	0.827 [0.799, 0.852]	0.858 [0.839, 0.876]
Paired contrast				
J4 vs J3 (primary, pre-registered)	+0.0033 [-0.0048, +0.0123]	-0.0004 [-0.0029, +0.0012]	+0.0000	-0.0015 [-0.0081, +0.0026]
J3 vs components (secondary)	+0.0395 [+0.0270, +0.0538]	+0.0055 [+0.0002, +0.0114]	+0.0258 [+0.0171, +0.0350]	+0.0344 [+0.0242, +0.0467]
Components vs released (convention)	+0.0734 [+0.0469, +0.1099]	+0.1601 [+0.1418, +0.1798]	+0.0560 [+0.0345, +0.0816]	+0.0568 [+0.0416, +0.0742]

Table 4: What the applied joins touch. Classes of the 20,578 J3 join rows against the site truth of the nine regions.

Class	Rows	Meaning
Joined, finished cell	238	same final root of a finished proofread cell: correct
Not joined, finished cell	31	different roots with a finished cell: wrong
Joined, unfinished cell	410	same root, proofread cell not finished
Joined, unproofread root	113	same root, nobody proofread it
Open	125	different roots, proofread but neither a finished cell
Not found	23	a side outside the site window
Neither side proofread	19,638	unverifiable (95.4%)
Decidable precision	0.885 [0.841, 0.918] (238 of 269); 0.571 with ambiguous rows counted wrong	

-0.001]); it failed its pre-specified no-material-cost rule and is not part of the pipeline. Development experiments on separate regions (Extended Data Table 2) found no resolved gain on the audited axon-like cohort at the fixed 1,000-row budget from more proofread cells within a region (30 to 260 cells) or from more training regions (4 to 24), although both raised all-cells run length and decidable precision at depth; no gain from 8 nm instead of 32 nm images once supervoxel shapes were in the input; and none from a second affinity recipe; and no automatic cutter reached an operating point above break-even. Joins can be learned from sparse truth; precise automatic cuts on axons were not achieved with this truth and these models.

Candidate repairs can be proposed without proofread labels

The pair model’s pre-training uses no proofreading: it learns from cuts made in the segmentation’s own segments, the construction of MergeNet¹⁷ run backwards: synthetic splits instead of synthetic merges. To ask whether that stage needs proofread labels at all, we trained the same three-dimensional network three ways and scored it unchanged on three held-out development regions holding 965 logged merges: on cuts only (no proofread label), on the 1,407 logged merges as positives, and on proofread-state labels (1,671 of 1,843 positives, the rest held out; Fig. 4, Table 5).

At 20,000 candidates per region the cuts-only network recovered 539 of the 965 logged merges, more than the proofread-state arm (472) on every region, with paired lower bounds above zero (+14, +15, +3), and more than the logged-merge arm on all three by point estimate (+10 [-6, +25], +21 [+6, +37], +5 [0, +10]). At 5,000 the paired intervals do not separate the arms; at 1,000, the budget the pipeline uses for its pair-model rows, the logged-merge arm is ahead by point estimate (152 against 138, intervals covering zero) and its purity among decidable pairs is lower. These single-seed comparisons used unequal training budgets, and the proofread-state arm’s negatives are mostly orphan attachments of unknown truth, which may handicap the supervised arms; they show useful proposal recall without proofread labels, not equivalence to supervised

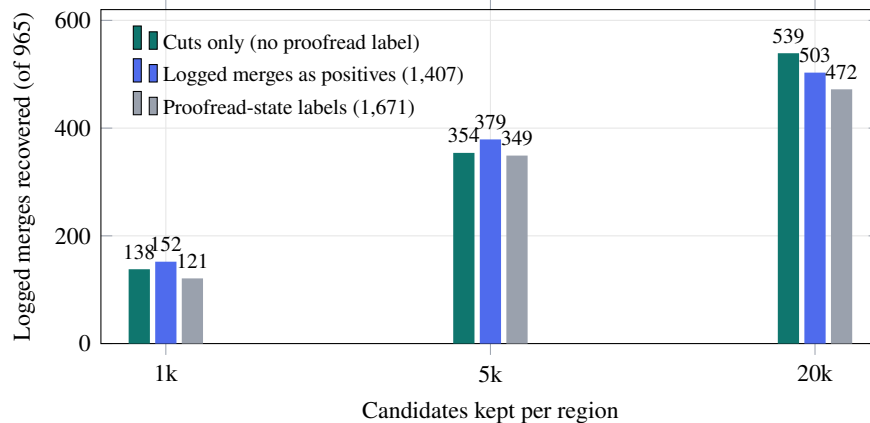


Figure 4: The candidate stage without proofread labels. Logged merges recovered among the top 1,000, 5,000 and 20,000 candidates per region, summed over three held-out development regions (965 logged merges), for the same three-dimensional network trained on cuts made in the segmentation’s own segments, on logged merges, or on proofread-state labels. One seed per arm. At 20,000 the cuts-only network recovers more than either labeled arm on every region by point estimate; its paired lower bound is above zero against the proofread-state arm in all three regions and against the logged-merge arm in one (at zero in a second); at 5,000 the paired intervals do not separate the arms; at 1,000 the logged-merge network is ahead by point estimate.

Table 5: Label-free candidate generation against the same network trained on real labels. Logged merges recovered per region at three budgets, and the paired difference of the cuts-only arm over the proofread-state arm (95% bootstrap intervals). One seed per arm; unequal training budgets. Dev. 1 to 3 are development regions, distinct from evaluation regions A to I.

Arm	Labels	Dev. 1 (430)	Dev. 2 (429)	Dev. 3 (106)	Sum (965)
Cuts only	0	62 / 161 / 243	66 / 153 / 233	10 / 40 / 63	138 / 354 / 539
Logged merges	1,407 merges	72 / 175 / 233	63 / 158 / 212	17 / 46 / 58	152 / 379 / 503
Proofread-state labels	1,671 of 1,843	55 / 163 / 213	51 / 142 / 204	15 / 44 / 55	121 / 349 / 472
Cuts only minus proofread-state, at 20,000		+30 [+14, +47]	+29 [+15, +45]	+8 [+3, +14]	

Counts are at the 1,000 / 5,000 / 20,000 budgets.

repair. The final pipeline fine-tunes on proofread labels.

Discussion

The results answer the four questions unevenly, which is the point of asking them separately. The proof-reading decision, as the benchmark poses it, is learnable to or beyond the published expert level by a small model, and its error falls with data at a rate (about 13% and 10% per doubling over the measured range) that is small per step and large over the decades of corrections that the existing reconstructions have already logged. Whether that translates into proofreading hours saved is not measured here. It is not, on its own, a finished brain. Transfer to expansion microscopy is partial but real on the barcode-readable subset of one region of one animal, and at the smallest label count the electron-microscopy checkpoint was worth four to eight times the barcode labels, which suggests the corrections of the old method can train the models for the new one. The end-to-end test shows that learned joins improve run length on annotated processes, in every region, while the image-based stage we expected to help did not meet its own criterion; we report the failed primary with the same prominence as the secondary success because the protocol was written to make that unavoidable.

The limitation that matters most is also the clearest. Run length is measured on proofread cells, because that is what the truth supports; 95.4% of the joins the pipeline applied connect tissue neither side of which has been proofread, and their precision is unmeasured. The decidable precision of 0.885 is a sample of 269 joins drawn from the 4.6% that touch proofread or partly proofread cells, and it may not describe the rest. An automatic repair that is right on the annotated minority and wrong on the unannotated majority would look like success under every metric in this paper and be worse than the released segmentation. The next measurement is therefore not a better pipeline but an independent annotation design that samples applied joins on currently unannotated tissue and scores repaired breaks and newly introduced wrong joins among them. Until that exists, these results support the use of learned joins as proposals for human review, with a measured precision on the finished-cell-decidable subset and an unknown precision on the rest, and not as unreviewed edits.

Three further limits bound the claims. The scaling and transfer results are on decision benchmarks whose task formats and label semantics come from the benchmark authors; “above human” is above a published expert point estimate on that format, not a head-to-head. The nine-region test is a held-out-region estimate within one heavily proofread data set, scored against its own proofreading, and says nothing about other animals or methods. And the development experiments report bounded negative findings under the tested budgets, architectures and supervision, not a general limit on what labels or images can give.

Methods

ConnectomeBench2 data

We downloaded a stratified sample of the ConnectomeBench2 v2 release⁹ from its public repository (sha256-verified): 100 of 669 training shards (79,500 decisions: 44,681 split correction with 15,682 positive; 34,819 merge identification with 16,047 positive), 20 of 105 validation shards (15,780) and 30 of 129 test shards (23,730: 12,849 split correction with 4,409 positive; 10,881 merge identification with 4,855 positive), and later all 903 shards (531,734 training decisions; the full test split of 101,929: 55,518 and 46,411). Every shard is stratified across the four species (mouse MICrONS, human H01, fly FlyWire, zebrafish Fish1) and five sample types, so a prefix of shards is a fair sample; the training sample held 33,590 mouse, 36,616 fly, 4,775 human and 4,519 zebrafish decisions. Decisions are real proofreader edits from the CAVE logs⁸ plus the authors’ negative controls, rendered at 224 pixels: three orthographic mesh views (seven channels) and three orthogonal electron-microscopy slices with segment masks, with the benchmark’s leakage rules (fused masks, no oblique slice for the single-segment task) and its 80- μ m cube-based split. Mouse and human rows are CC BY 4.0 upstream; fly and zebrafish CC BY-NC.

Model and training

Two ImageNet-pretrained ConvNeXt backbones¹⁴ from the timm library, one shared across the three mesh views and one across the three image slices, view features averaged within each stream, concatenated, and passed to a multilayer perceptron with one logit per task. Sizes: ConvNeXt-Atto (6.9 million parameters in total) and ConvNeXt-T (56.0 million); for Extended Data Table 1 also ConvNeXt-S, ViT-B/16 (two-stream, 172 million), a DINOv3-pretrained¹⁵ set of the same, and an electron-microscopy-pretrained ViT-L/14. AdamW, weight decay 0.05, batch 64, 3% warm-up then cosine decay, gradient clipping at 1, stochastic depth 0.1, per-task positive-class weights, light augmentation (per-view flips, image intensity jitter). Learning rates: the Atto sweep and its exponents at 5×10^{-4} , the grid value (a separate learning-rate sweep at 79,500 decisions favored 2×10^{-4} , which gained 0.6 and 3.2 percentage points there); the Atto full-data runs at 5×10^{-4} (a 2×10^{-4} full-data run traded 0.19 points of split correction for 0.31 of merge identification); ConvNeXt-T at 5×10^{-5} in the sweep and 2×10^{-4} in the three-seed full-data runs; the DINOv3 ConvNeXt-T at 2×10^{-5} . Sweep runs targeted 120,000 training examples with the epoch count clamped to 3 to 40, so the examples actually seen were 38,400, 119,040, 128,000 and 238,464 at the four sizes; data and optimization budget

therefore vary together. Full-data runs trained for ten epochs. For sweep runs the checkpoint with the best mean balanced accuracy on a 2,000-example validation subset was scored once on the held-out test decisions; for full-data runs the final (epoch 10) checkpoint was scored and the best validation epoch (6 to 10) is recorded per run. The test decisions were never used for training or selection. Sweep runs were on an Apple M-series GPU, NVIDIA H100s and an RTX 5090 (the tuned ConvNeXt-T points); full-data runs on H100s (100 to 110 minutes for Atto, 221 to 478 for ConvNeXt-T).

Scaling fits and comparisons

Error (one minus balanced accuracy) against training decisions was fitted as $aN^{-\alpha}$ by least squares on the three-seed means at 1,000, 4,000, 16,000 and 79,500 decisions; standard errors are least-squares, bootstrap intervals resample each point inside its test-set interval. The full-data points were not used in the fits; their deviation from the fit’s extrapolation is reported. Balanced-accuracy intervals on single runs are 95% bootstraps over test rows. Human-expert figures and the published ViT-B figures are the benchmark preprint’s⁹; the published model is a single Vision Transformer with two patch embeddings sharing one encoder, trained from ImageNet weights on the benchmark’s 75% training split and scored on its held-out test set, so comparisons with it are on the same task format and about the same training data; whether its test items coincide with our 101,929-decision split is not established.

PRISM labels, rendering and transfer

E11 Bio’s evaluation region (mip 1: 336×336×400 nm voxels in expanded space, about 70×70×80 nm native at 4.8× lateral and 5.0× axial expansion; 3,380×3,380×366 voxels)¹². Decisions were made on E11’s GFP-only segmentation (896,322 segments). Barcode readout per segment and per 48×48×40-voxel part from E11’s 18-channel bit-intensity estimate, normalized against a baseline (median over the 9,684 segments of at least 2,000 voxels) and a robust scale (1.4826× lower-half median absolute deviation, floor 0.5); a segment is readable with at least 300 voxels, vector norm at least 4 and at least 90% of its coded parts within cosine 0.9 of the segment vector (73.4% of the 48,815 segments of at least 300 voxels). Split correction: a face-touching pair (at least 8 faces, both at least 300 voxels) is labeled same-cell if both are readable with cosine at least 0.9, identical binary code and not a collision code (a code shared by two or more skeleton-traced cells), and different-cell if cosine at most 0.3 with Hamming distance at least 2. Merge identification: a segment of at least 2,000 voxels contains a merge if its coded parts split into two groups with cosine at most 0.3 and Hamming distance at least 2, with a minority of at least 200 voxels and 5%, and is one cell if readable, not a collision code, with at least three coded parts covering at least half its volume. The labels cover about 13% of candidate pairs, from one animal and one region (CA3, about 0.19 mm³ expanded). Rule thresholds were chosen on error rates against random pairs (the same-cell rule fires on 9.9×10^{-5} of random different-cell pairs; the different-cell rule on 2.6% of same-cell pairs), not on model results. Against the human skeletons where both exist, barcode labels agree on 104 of 105 split-correction and 73 of 79 merge-identification decisions. Examples were rendered in the ConnectomeBench2 format from the GFP channel and the segment meshes at the electron-microscopy fields of view scaled by the expansion factor; no barcode channel enters any input. Train, validation and test regions were separated along one axis with the test region beginning 450 voxels (about 31 μ m native) beyond the training region; decisions crossing a boundary were dropped, as were 404 test decisions touching a skeleton cell also touched by a training decision. Fine-tuning: ConvNeXt-Atto at fixed compute per run (40,000 examples seen, batch 64), ImageNet initialization at learning rate 5×10^{-4} or the electron-microscopy checkpoint at 1×10^{-4} (each start at its own rate, so initialization and learning rate are not separated), seeds 0, 1 and 2 at each of 250, 500, 1,000, 2,000 and 3,865 labeled decisions (30 runs); the seed changes the subset, the head initialization and the data order. The score is AUROC on the held-out test region; power laws on the error were fitted to the five seed means with bootstrap intervals from 2,000 draws.

The nine-region evaluation

Regions. Nine regions of minnie65, chosen by a written rule from a ranking of candidate boxes by logged-edit density, never used for training, tuning or selection, at least 30 μm from every region that trained a model used on them (one excluded model family, trained on a region 25 μm away, was not used), sealed until the protocol was frozen; 21 to 50 mm of preprint-rule reference path each (24 to 57 mm on the standing skeletons). Four further regions, sealed with a 15 μm clearance, remain unopened. *Pipeline.* (1) Connected components within each padded evaluation crop. (2) The pair model: a three-dimensional network pre-trained on 14,268 cuts made in the segments of one development region (20,000 steps, no proofread label), then fine-tuned on 1,294 positive join examples labeled from the proofread state of four training regions (4,000 steps); candidate joins scored over supervoxel pairs, the top 1,000 rows per region applied. (3) The tip judge: a boosted-tree classifier of “does this tip continue?” retrained deterministically from the four training regions’ tip tables, threshold 0.96, its rows added to the join list. States: $J3 = (1)+(2)+(3)$; $J4 = J3$ plus image-based joins from a dense affinity network on $16 \times 16 \times 40$ nm voxels, applied where the pooled affinity is at least 0.9 over at least 4,096 voxel faces. A label-free cleaner (supervoxel-shape classifier at threshold 0.98, segments of at least 200 supervoxels eligible) was evaluated as a secondary ablation. *Pre-registration.* The candidate ($J4$), the comparator ($J3$), the success rule (gain at least 0.005 in plain nERL and paired 95% lower bound above zero), the evaluators and the sha256 hashes of 72 artifacts (checkpoints, scripts, inputs) were written to a GO file before the regions were opened; an execution record of 291 lines logged the hash of every produced file and the machine that produced it. On reconciliation, two small source files in the hash list did not match the versions run: a pod launcher edited three times under the protocol (fetching the GO file, judging success by the saved score file, extending the hash line) and a join script with three changes to output paths and a summary print, none to a computed value; both corrections are in the record with their timestamps. Candidate-side steps read a label-free cache; the truth was read only by the evaluator. Five protocol amendments made at execution (a machine substitution after a GPU failure, the checkpoint-loading path of record for the image model, the cleaner’s launcher, the cleaner’s threshold, one fallback placement) are recorded with their timestamps. *Metrics.* Normalized expected run length on TEASAR skeletons¹⁸ of the proofread neurons: plain nERL, strict nERL (the same with any wrong join zeroing the neuron), the audited axon-like cohort (82.8 mm classified axon-like by synapse counts) with wrong joins detected against every proofread skeleton, and the all-cells audited evaluator, all on the preprint-rule reference skeletons (TEASAR scale 5, constant 0, 100 nm endpoint erosion, 300 nm resampling; 345.5 mm). Breaks and wrong joins per millimeter by minimum-overlap counting on our standing skeletons (TEASAR scale 1.5, constant 300 nm; 388.9 mm); the two are skeletonizations of the same proofread cells. Intervals are 95% paired bootstraps grouped by neuron across regions; leave-one-region-out ranges and the concentration of the gain over neurons are reported for the primary. Decidable join precision uses the finished-cell site truth: a join is correct if both sides belong to the same final root of a finished proofread cell and wrong if they belong to different roots with a finished cell; its interval is a Wilson interval treating joins as independent.

Label-free candidate stage

The same three-dimensional network trained three ways: on cuts only (as above), from scratch on the proofread-state labels of the four training regions (1,671 of 1,843 positives, the other 172 held out; 1,562,942 negatives, 8,000 steps), and on the four regions’ 1,407 logged merges as positives against all other rows. Scored unchanged on three held-out development regions (430, 429 and 106 logged merges) as the number of logged merges among the top 1,000, 5,000 and 20,000 candidates per region, with paired bootstrap intervals on the differences (10,000 draws for the cuts-only minus logged-merge contrast). One seed per arm.

Reproducibility

The pair model’s validation scores reproduced bit for bit on the development GPU after a driver update, and to identical top-1,000 membership on a second GPU type (differences within half-precision arithmetic, at most 0.0156 and 0.0313 in score for the two models); the join chain and the shape scorer reproduced byte for byte from the label-free cache; three fresh-process reruns of the image model on three GPU types agreed to about 0.1 of 255 and gave the identical J4 state (41 joins on the development region), while the original in-process validation table differed by 2.4 of 255 (39 joins, 38 shared) with the run-length rows unchanged to three decimals; rebuilding J3 and J4 from GPU-noisy base scores changed no decision. An eight-round review by a separate AI model, reading only the saved per-neuron outputs, reproduced the intervals. Operational deviations (a GPU failure mid-run, three lost pod launches, two path fixes at execution) are documented with the replay checks.

Statistics

Balanced accuracy is the mean of sensitivity and specificity at a 0.5 threshold; AUROC is the area under the receiver operating characteristic; nERL as defined above. Uncertainty is reported as: 95% bootstrap intervals over test rows for single benchmark runs; the seed standard deviation for sweeps; least-squares standard errors and bootstrap intervals for power-law exponents; 95% bootstrap intervals over neurons grouped across regions for the end-to-end test, paired for contrasts; a Wilson interval for decidable join precision; and paired t intervals over six region-by-seed differences for the 8 versus 32 nm comparison. Contrasts other than the pre-registered primary are secondary or exploratory, and no multiplicity correction was applied.

Use of AI tools

AI-assisted tools were used for code generation, for bookkeeping of the runs and for checking the analysis and wording; every reported number derives from the logged runs and their hashed outputs, and the author takes responsibility for the content.

Data availability

ConnectomeBench2 is public⁹. The MICrONS minnie65 segmentation and proofreading are public through the MICrONS program^{3,8}. E11 Bio’s PRISM data are public (CC BY-SA 4.0 with a separate restriction on commercial use; research use here)¹². Our per-run tables (every sweep and full-data run with its command line), the scaling and transfer fits, the GO file (sha256 174b5ca6...), the execution record, the per-region and per-neuron run-length outputs and the report tables are archived with the author and available on request.

Code availability

Training, rendering, labeling, pipeline and scoring code is archived with the author and available on request.

Acknowledgments

We thank the ConnectomeBench2 authors for the benchmark and its expert figures, the MICrONS and FlyWire consortia for public reconstructions and proofreading logs, and E11 Bio for the PRISM release.

Competing interests

D.S. is the co-founder and CEO of Big Brain Lab, which develops brain-mapping methods and would benefit from automated proofreading.

References

- [1] Dorkenwald, S. *et al.* Neuronal wiring diagram of an adult brain. *Nature* **634**, 124–138 (2024).
- [2] Schlegel, P. *et al.* Whole-brain annotation and multi-connectome cell typing of *Drosophila*. *Nature* **634**, 139–152 (2024).
- [3] The MICrONS Consortium. Functional connectomics spanning multiple areas of mouse visual cortex. *Nature* **640**, 435–447 (2025).
- [4] Shapson-Coe, A. *et al.* A petavoxel fragment of human cerebral cortex reconstructed at nanoscale resolution. *Science* **384**, eadk4858 (2024).
- [5] Berg, S. *et al.* Sexual dimorphism in the complete *Drosophila* male central nervous system connectome. *Cell* **189**, 5504–5526.e15 (2026).
- [6] Januszewski, M. *et al.* High-precision automated reconstruction of neurons with flood-filling networks. *Nat. Methods* **15**, 605–610 (2018).
- [7] Wellcome. *Scaling up connectomics: the road to a whole mouse brain connectome* (Wellcome, London, 2023). https://cms.wellcome.org/sites/default/files/2023-06/Connectomics-scaling-up-connectomics_0.pdf
- [8] Dorkenwald, S. *et al.* CAVE: Connectome Annotation Versioning Engine. *Nat. Methods* **22**, 1112–1120 (2025).
- [9] Brown, J., Farkas, T., Razgar, G. & Boyden, E. S. ConnectomeBench2: a unified benchmark for automated connectomic proofreading. *arXiv* 2606.21116 (2026).
- [10] Tavakoli, M. R. *et al.* Light-microscopy-based connectomic reconstruction of mammalian brain tissue. *Nature* **642**, 398–410 (2025).
- [11] Glaser, A. *et al.* Expansion-assisted selective plane illumination microscopy for nanoscale imaging of centimeter-scale tissues. *eLife* **12**, RP91979 (2025).
- [12] E11 Bio. PRISM: protein barcodes for expansion-microscopy connectomics. <https://www.e11.bio/blog/prism-and-the-public-data-release-s3://e11bio-prism> (2025).
- [13] Januszewski, M. *et al.* Accelerating neuron reconstruction with PATHFINDER. *bioRxiv* 10.1101/2025.05.16.654254 (2025).
- [14] Liu, Z. *et al.* A ConvNet for the 2020s. *Proc. CVPR* (2022); *arXiv* 2201.03545.
- [15] Siméoni, O. *et al.* DINOv3. *arXiv* 2508.10104 (2025).
- [16] Brown, J., Kirjner, A., Vivekananthan, A. & Boyden, E. ConnectomeBench: can LLMs proofread the connectome? *arXiv* 2511.05542; *NeurIPS Datasets and Benchmarks Track* (2025).
- [17] Rolnick, D. *et al.* Morphological error detection in 3D segmentations. *arXiv* 1705.10882 (2017).
- [18] Sato, M. *et al.* TEASAR: tree-structure extraction algorithm for accurate and robust skeletons. *Proc. Pacific Graphics* (2000).

Extended Data

Extended Data Table 1. Backbones at 79,500 training decisions. Balanced accuracy on the 23,730 held-out decisions, three epochs, one seed each unless stated; two-stream models throughout; minutes are H100 wall clock, blank for the two RTX 5090 runs. The learning rate mattered more than the backbone family.

Backbone (two streams)	Pretraining	Learning rate	Split corr.	Merge ident.	H100 minutes
ConvNeXt-S, 100M	DINOv3	2×10^{-5}	96.9	93.6	16
ConvNeXt-T, 56M	ImageNet	5×10^{-5}	96.9	93.5	(RTX 5090)
ConvNeXt-T, 56M (5 epochs)	ImageNet	5×10^{-5}	96.8	93.8	17
ConvNeXt-T, 56M (5 epochs)	DINOv3	2×10^{-5}	96.8	93.5	17
ViT-B/16, 172M	DINOv3	1×10^{-4}	96.6	92.8	18
ConvNeXt-T, 56M	ImageNet	1×10^{-4}	96.6	93.1	(RTX 5090)
ConvNeXt-T, 56M	DINOv3	2×10^{-5}	96.5	92.7	10
ConvNeXt-T, 56M	ImageNet	2×10^{-4}	96.3	91.7	10
ConvNeXt-Atto, 6.9M	ImageNet	2×10^{-4}	96.3	92.2	4.6
EM-DINO ViT-L/14, 608M	EM-DINOv2	2×10^{-5}	96.4	92.4	62
ConvNeXt-Atto, 6.9M (3 seeds)	ImageNet	5×10^{-4}	95.7	89.0	4.6
ConvNeXt-T, 56M	DINOv3	1×10^{-4}	94.8	87.0	10
Full data, 531,734 decisions, full test split (101,929)					
ConvNeXt-Atto, 3 seeds	ImageNet	5×10^{-4}	97.41 (0.04)	94.38 (0.06)	100–110
ConvNeXt-Atto, 1 seed	ImageNet	2×10^{-4}	97.22	94.69	101
ConvNeXt-T, 3 seeds	ImageNet	2×10^{-4}	97.54 (0.10)	94.82 (0.12)	221–478
ConvNeXt-T, 1 seed	ImageNet	5×10^{-5}	97.68	95.32	224
ConvNeXt-T, 3 seeds	DINOv3	2×10^{-5}	97.75 (0.06)	95.52 (0.05)	235–278
ViT-B/16, 1 seed, two encoders	ImageNet	1×10^{-4}	97.29	93.90	264
ViT-B/16, 1 seed	DINOv3	1×10^{-4}	97.40	94.98	408
ViT-B/16, 1 seed	DINOv3	2×10^{-5}	97.63	94.94	393
Published ViT-B, shared encoder, training split	ImageNet		97.0	93.0	about 900 GPU-min

Extended Data Table 2. Development experiments that did not help. Bounded negative findings under the tested budgets, architectures and supervision, on development regions with seeds and paired intervals (the cleaner row is from the nine-region evaluation); not a general limit.

Lever	Result
Proofread cells within one region (30 / 100 / 260 cells; 5 draws or 3 seeds)	Audited axon-like 0.807 / 0.806 / 0.803 against baseline 0.798 (draw s.d. 0.006 / 0.001 / 0.001); all-cells plain rose 0.846 / 0.854 / 0.867.
Training regions (4 / 8 / 16 / 24; 3 seeds)	Audited axon-like 0.818 / 0.817 / 0.817 / 0.819 against baseline 0.803; paired bounds under +0.012.
The 16-region model inside the full pipeline	Against J3: plain -0.021 / -0.005 at budgets 1,000 / 2,000; its extra joins at decidable precision 0.71 to 0.79.
8 nm instead of 32 nm images (2 regions, 3 seeds; paired t intervals)	Boundary AUROC -0.0004 [-0.005 , $+0.004$] with supervoxel shapes in the input; $+0.007$ [$+0.001$, $+0.013$] image only.
Automatic cutters	2 of 50 applied cuts correct; no operating point clears break-even; no threshold passes the precision bar.
Second affinity recipe (7 regions, 2 seeds)	Cut AUROC $+0.01$ to $+0.08$, join AUROC unchanged; no gain for the image-join rule.
Generated negatives (5 seeds)	Shape-checker AUROC fell from 0.790 to 0.775 under the tested recipe.
Label-free cleaner before J3	Strict $+0.017$ [$+0.011$, $+0.024$], audited axon-like -0.008 [-0.016 , -0.001]; failed its no-material-cost rule.

Extended Data Table 3. Per-region result of the primary comparison. Plain nERL under J3 and J4 and their paired difference, per evaluation region; reference path on the preprint-rule skeletons.

Region	Reference path (mm)	J3	J4	J4 minus J3
A	35.6	0.7562	0.7580	+0.0018 [-0.0001, +0.0046]
B	45.4	0.8989	0.9003	+0.0014 [+0.0000, +0.0047]
C	21.4	0.6487	0.6715	+0.0228 [-0.0003, +0.0685]
D	41.9	0.8701	0.8711	+0.0010 [+0.0000, +0.0034]
E	49.7	0.8714	0.8571	-0.0143 [-0.0516, +0.0073]
F	49.9	0.8945	0.8969	+0.0024 [+0.0000, +0.0080]
G	27.6	0.9322	0.9322	+0.0000
H	24.3	0.8965	0.9496	+0.0531 [+0.0000, +0.1508]
I	49.7	0.8588	0.8594	+0.0006 [+0.0000, +0.0018]
Pooled	345.5	0.8575	0.8608	+0.0033 [-0.0048, +0.0123]