High-accuracy neurite reconstruction for high-throughput neuroanatomy

Moritz Helmstaedter, Kevin L Briggman, Winfried Denk

High-accuracy neurite reconstruction for high-throughput neuroanatomy.

Moritz Helmstaedter, Kevin L Briggman, and Winfried Denk

Max Planck Institute for Medical Research, Jahnstr. 29, D-69120 Heidelberg, Germany.

Running title: High-accuracy 3D EM skeletonization

Key words: serial block-face electron microscopy, neural circuit reconstruction, connectomics

Editorial correspondence: Moritz Helmstaedter, Max Planck Institute for Medical Research, Jahnstr. 29, D-69120 Heidelberg, Germany. Phone: +49 6221 486 149. Fax: +49 6221 486 325; E-mail: moritz.helmstaedter@mpimf-heidelberg.mpg.de

Competing financial interests: Published Patent Application US 20100183217 (MH and WD). IP License income from Gatan Inc. for Serial Blockface Imaging (WD).

ABSTRACT

Neuroanatomic analysis depends on the reconstruction of complete cell shapes. High-throughput reconstruction of neural circuits ("connectomics") using volume electron microscopy requires dense staining of all cells, where even experts make annotation errors. Currently, reconstruction rather than acquisition speed limits the determination of neural wiring diagrams. We present methods for the fast and reliable reconstruction of densely labeled datasets. Our approach, based on manually skeletonizing each neurite redundantly (multiple times) with a special visualization/annotation software tool (KNOSSOS), is ~50 times faster than volume labeling. Errors are detected and eliminated by a "redundant-skeleton consensus procedure" (RESCOP), which uses a statistical model of how true neurite connectivity is transformed into annotation decisions. RESCOP also estimates the consensus skeletons' reliability. Focused re-annotation of difficult locations promises a rather steep increase of reliability as a function of the average skeleton redundancy and thus the nearly error-free analysis of large neuroanatomical datasets.

INTRODUCTION

The reconstruction of neuronal circuits has been a central approach toward understanding the function of the nervous system since the earliest studies by Golgi and Ramòn y Cajal[1-2]. While many neurons extend over tens of centimeters, the caliber of thin neurites can be as small as 40 nm (spine necks[3]). This range of length scales is bound to challenge any method aimed at the extraction of neuron morphology from the data. For sparsely stained tissue, with only a small fraction of all neurons labeled, such as with the Golgi method[2] or by selective dye injection[4-5], imaging techniques operating at a resolution of around 1 μm are sufficient to follow all processes. This holds true even if the neurite caliber falls well below the imaging resolution, because in very sparsely stained data the identity of each neurite is easily established. Manual reconstructions of individual neurons from such data are, therefore, assumed to be highly reliable, even though little validation of this reliability has been reported. Almost all available neuroanatomical data at single-cell resolution stem from such experiments, but as fluorescence imaging data from samples with a much higher staining density are becoming available (hundreds of neurons per 1 mm^3, labeled using various genetic or virus-based techniques[6-7]), high reconstruction reliability can no longer be presumed.

For the reconstruction of complete cellular wiring diagrams ("connectomes"[8-9]) assuring reconstruction reliability is even more difficult because the morphologies of all neurons, not only those of a small subset, have to be extracted. This may eventually be possible at light-microscopic resolution by staining all neurons with a sufficient number of distinguishable colors[7, 9] but otherwise requires imaging at a resolution high enough to follow all neurites in densely packed neuropil (discussed in[10]). Such a reconstruction was performed for the entire nervous system (302 neurons) of the nematode *C. elegans*[11] using serial-section electron microscopy.

Recently developed techniques for automated volume electron microscopy[12-15] enable the imaging of volumes large enough to contain more complex neural circuits[16]. However, extracting information about neuron morphology and circuit structure from such data poses two major challenges. First, the total neurite path length in many neural circuits is typically in the range of meters (at least 0.3 m for small circuits such as a $(100\ \mu m)^3$ region of retina, and as much as 400 m for a mouse cortical column[10]). Using currently available software tools for neurite contouring (e.g., Reconstruct[17]) the complete analysis of such circuits is very slow and thus prohibitively expensive. Contouring every neurite for a path length of 0.3 m would require an estimated 60,000 hours (30 person years) of annotation time. Reconstruction *accuracy* is the second major concern. While for sparsely stained data the selectivity of the stain makes following the neurites easy, connectomic reconstruction requires a large number of decisions (as many as one every ~4 μm in the retina) about whether to continue, branch, or terminate a neurite. Some of these decisions are difficult and, more importantly, because they have to be made constantly while annotating, their reliability depends on the uninterrupted attentiveness of the human annotator. As a third obstacle, synapses must be identified with sufficient accuracy.

Here, we describe a set of tools that substantially improve both the speed and the accuracy of neurite reconstruction. We chose to annotate the data by following a single core line along the inside of each neurite, creating a "skeleton" representation of each neuron's morphology. When using the KNOSSOS software tool, which we developed for the convenient browsing and annotation of large datasets, we observed a 50-fold (range 20-130-fold) increase in the amount of neurite path length reconstructed per unit time. We quantified discrepancies between multiple (redundant) skeletons of the same neurite and, based on their distribution, optimized the correction of errors and the creation of a consensus skeleton (which is actually a bundle of closely spaced skeleton pieces). We call our method "REdundant-Skeleton COnsensus Procedure, RESCOP", with 'redundant' used

in the sense of “multiple redundant systems working in parallel to increase reliability”. We show that the accuracy of the consensus skeleton quickly rises with the number of redundant skeletons, even when employing only slightly trained annotators. We have used our set of reconstruction tools to skeletonize all rod bipolar cells in a block of mouse retina.

RESULTS

Browsing large-scale EM data

We first developed a software tool (KNOSSOS, s. Supplementary Movie) for browsing and annotating large-scale volume data. Such data are generated, for example, by serial block-face electron microscopy (SBEM[12]). At nanometer resolution, imaging volumes that are large enough to contain entire circuits yields datasets that are at least several hundreds of gigabytes in size. KNOSSOS was designed to make 3D navigating and viewing of such data sets convenient. KNOSSOS allows quick navigation along all axes by selectively loading only the data surrounding the currently viewed location. Neurites can be oriented along any direction in dense neuropil and can often be followed more conveniently using views other than the imaging plane (the block face in SBEM), in particular if the data, as is the case for SBEM, are nearly isotropic in resolution. KNOSSOS, therefore, displays three orthogonal views of the data (see also V3D[18]), which were found to be essential to navigate along neurites oriented obliquely to the slice plane. KNOSSOS runs smoothly on laptops with the data located on an external hard drive. This allowed us to distribute the work load to a large number of non-expert annotators (in our case >80 undergraduate students).

Fast neurite reconstruction by skeletonization

In order to densely reconstruct even a local neuronal circuit, at least several hundred millimeters of neurite need to be correctly followed. This can, in principle, be done by contouring (volume labeling) of neurites (**Fig. 1a**). However, contouring is slow (200-400 hours per mm of neurite length[10]).

KNOSSOS, therefore, provides a skeletonization mode (**Fig. 1b**, **Supplementary Movie**). The user starts at a location within a neuron (which we call "seed"), for example the cell body, and places a marker (which we call "node", **Fig. 1c**). Then, the user advances through the data along a neurite, and places nodes at intervals of approximately 7-10 image

planes, approximately at the center of the neurite. Importantly, the user can move in any of the cardinal directions, and can place nodes in any of the three orthogonal view ports. Sequentially placed nodes are connected by line segments (which we call "edges", **Fig. 1c**). When a location where the neurite branches is encountered, the user designates the current node as a branch point, and is later directed back to this branch point after completing one of the branches. Skeletonization allows the user to focus annotation to the core line of a neurite. We found that skeletonization reduced annotation time to 5.9±2.8 hours per millimeter path length, which is roughly 50-fold (range 20-130-fold) faster than fully manual volume labeling (**Fig. 1d**, **Supplementary Fig. 2**).

Discrepancies between skeletons

We next investigated how frequently annotators disagreed when skeletonizing the same neurite, starting from the same initial location. **Figure 2** shows the overlay of 2 skeletons generated by two experts (i.e., experienced neuroscientists), both starting at the soma of an amacrine cell in a SBEM dataset of rabbit retina (dataset E1088, s. Methods). The skeletons disagreed at 12 locations along the dendritic tree, which has a total path length of 0.8 mm. Most of the disagreements (10 out of 12) were caused by missed branch points (**Fig. 2**, locations 1, 2, 4, 6-12, see **Supplemental Material** for image stacks centered at those disagreement locations). Upon re-inspecting these 10 locations, both annotators quickly reached agreement, which suggests that the missed branch points had simply been overlooked. This implies that continuous attention is needed in order not to miss any of the branches along the neurite. Two of the disagreements (locations 3, 5) were not missed branches but instead were locations where one annotator continued the neurite skeleton and the other annotator did not. While one of these two locations (location 3) was easily resolved, agreement between the annotators could be reached for location 5 only upon close inspection, which means this location was truly difficult to annotate. In this case, the difficulty was caused by the local neurite geometry (a tip-to-tip contact). We similarly

found both attention- and difficulty-related errors when skeletonizing axons in fluorescent data (imaged by confocal microscopy; data not shown). It is this variation in difficulty that can be captured by our statistical model of neurite detectability, introduced below.

These initial results indicated that even experts make annotation errors and that skeletons have to be cross-checked. We, therefore, proceeded to further quantify skeleton accuracy across a number of annotators, and then developed an algorithm to find the consensus skeleton and to estimate its accuracy.

Error quantification

To detect errors in the skeletons, we asked multiple annotators to skeletonize the same neurite (**Fig. 3a**). For each edge (line segment, s. above), that one of the annotators had created, we then measured how many of the other annotators agreed with the decision to create this edge (**Fig. 3b**). Our agreement measure is based on the following reasoning: when one annotator skeletonized an edge he/she made the decision that the neurite continues at the location of this edge. A second annotator *agreed* with this decision if his/her skeleton also reaches the edge location *and* continues beyond it. Conversely, a second annotator *dis*agreed with this decision if his/her skeleton reaches this location but does *not* continue. To detect and distinguish these two cases we used the following procedure to evaluate the proximity between skeletons.

To evaluate an edge created by one of the annotators we first considered only the edge in question plus a few edges on each side (skeleton A, **Fig. 3b**), yielding an evaluation spotlight moving along the skeleton (**Fig. 3b**). The size of the spotlight was a sphere of on average 700 nm radius, depending on how closely the annotator had placed the neighboring skeleton nodes. (**Fig. 3b**; for the choice of the spotlight size s. below and Methods). We then temporarily removed the edge in question, splitting the skeleton into two pieces (**Fig. 3c**), and then measured the distances between each of these two skeleton pieces and all the other annotators' skeletons (skeletons B, C, D, etc). If another annotator's skeleton

(skeleton C in **Fig. 3b,d**) was close enough to both skeleton pieces, this other annotator was considered to have voted for the edge in question (agreeing vote, **Fig. 3d**). Conversely, if another annotator's skeleton was close to only one of the skeleton pieces (skeleton D in **Fig. 3b,e**), this other annotator was considered to have voted against the edge in question (disagreeing vote, **Fig. 3e**), because this corresponds to a skeleton reaching the location of the edge but *not* continuing. If the other skeleton was too distant from both skeleton pieces it likely belonged to a different neurite and was, therefore, disregarded. Skeletons were considered close enough when the root-mean-square distance between the nodes of the skeleton piece and the edges of the other annotator's skeleton was smaller than 625 nm. The value of this maximal distance, and the value of the spotlight radius used above were determined by searching for those parameters that minimized the disagreements between a 50-fold and 15-fold consensus skeletons (s. below and Methods). Note that this procedure for measuring the agreement between skeletons requires a sufficient node density but does not require the node density to be the same or the node locations to be in register for different skeletons.

After applying this distance measurement to all edges in all annotator's skeletons, we obtained for each edge the number of agreeing votes and the total number of votes cast for that particular edge (the sum of agreeing and disagreeing annotators). We then counted the number of edges that had a certain combination of agreeing and total votes (say, 6 agreeing votes out of 10 total votes), and reported these for all encountered combinations of agreeing and total votes in a 2-dimensional vote histogram (**Fig. 3f**).

The distribution of inter-annotator agreement

We had one amacrine cell (~600 μm total neurite path length) skeletonized by 50 different annotators (s. **Fig. 4e**, left). Before voting we divided the set of 50 skeletons 3 times into two subsets, to which skeletons were randomly assigned. This created 6 subsets of 25 skeletons each. Their vote histograms were calculated separately, in order to later assess

the variability of our procedure, but for now we used the sum of these vote histograms (**Fig. 3f**, left panel). Note that most parts of the amacrine cell were found and annotated by all (25 total votes) or almost all (~20-24 total votes) annotators (**Figs. 3f, 4e**). Since some branches were followed only by a few annotators the vote histograms also contains entries for small number of total votes (**Fig. 3f**). In this histogram, we found complete agreement between annotators (number of agreeing votes equal to the total number of votes, evaluated for edges with at least three votes) for 68 % of all locations, for 8 % only one annotator disagreed, and 10 % of the locations were annotated by only one annotator. The locations where one annotator disagreed can be interpreted, at least for a large number of total votes, as having been missed due to inattention. The locations found by only one annotator were interpreted as erroneous continuations or branches. Most of the remaining 14 % of locations, where more than one annotator disagreed, are presumably more difficult locations in the data, because it is unlikely that two or more attention-related mistakes occur at the same location.

To measure annotation agreement for different kinds of neurites from different types of cells, we also calculated the vote histogram (**Fig. 3f**, right panel) for 98 skeletonized neurite fragments densely packed in another region of the same data set (**Supplementary Fig. 4**, 166,472 annotated edges with a total path length of 43.2 mm). In this case the total number of votes was lower on average (3.2±2.9, **Fig. 3f**, right panel) and varied much more. In both cases most annotators agreed for most edges, i.e. the votes were concentrated near the diagonal of the vote histogram. The vote histograms can be used to compare the difficulty of datasets, provided that the annotators were similarly trained and similarly attentive.

Skeleton consensus rules

Our next goal was to find the consensus skeleton based on multiple annotations of the same neurite by eliminating edges that were unlikely to be correct, based on the number of

agreeing and disagreeing votes. The intuitive choice for whether to accept or eliminate an edge is the majority vote, but it is not clear whether this is also the optimal decision. We, therefore, analyzed the annotation process in order to determine the rule to find the best consensus skeleton and to estimate the residual error rate of the consensus skeleton.

The model for annotation decisions

To describe the annotation process we used the following decision model, which reflects the fact that the annotation difficulty varies with location (**Fig. 3i,k**). While two intracellular voxels are either connected (i.e. belong to the same neurite) or not connected (i.e. belong to different neurites), this ground truth is to some degree obscured by fixation, staining, and imaging of the sample at limited resolution and signal-to-noise ratio. This makes annotation an inherently noisy process, with a probability, p_e, for each pair of points that annotators will create an edge, i.e. label the points as connected (**Fig. 3i,k**, middle, we also refer to p_e as edge detectability). The edge detectability depends on whether the points are actually connected (see below), but it also varies as a consequence of the local neurite geometry (wide, straight, or bundled neurites are, for example, easier to follow) and local staining quality.

In this model, the decision to create an edge between a pair of points corresponds to a biased coin toss, with the bias equal to the edge detectability p_e. Therefore, the decisions of the annotators will follow binomial statistics with a bias of p_e (**Fig. 3i**, bottom, Methods Eqn. 4). Obvious neurite continuities (where the edge detectability is close to 1, $p_e \approx 1$) and neurite discontinuities (where the edge detectability is close to 0, $p_e \approx 0$) will both result in a high agreement amongst annotators. Difficult locations have edge detectabilities p_e closer to 0.5.

We cannot determine the edge detectability, p_e, at a given location directly (except by annotating it a very large number of times). However, for any assumed distribution of edge detectabilities $p(p_e)$ in the data, we can compute the expected distribution of agreeing and

disagreeing votes (predicted vote histograms, for details see Eqns. 5,7, Methods). Comparing measured and predicted vote histograms (**Fig. 3f** vs. **g**), allowed us to search for the optimal distribution of edge detectabilities p(p_e), i.e. that best explained the measurements. We found that the optimal distribution of edge detectabilities p(p_e) consists of a number of peaks with a large peak near one (**Fig. 3h**), which reflects the high frequency of obvious neurite continuities. Note that because we cannot measure zero total votes the fit is not well constrained near p_e=0. In fact, a delta function at p_e=0 can be added to the distribution of edge detectabilities p(p_e) without changing the goodness of the fit and without affecting the following results.

To explore how variable the distribution of edge detectabilities p(p_e) is for different annotations of the same cell, we separately fitted vote histograms for the 6 sets of 25 out of 50 skeletons and found similar distributions of edge detectabilities p(p_e) (**Supplementary Fig. 5c,d**); What varies is the exact location of the peaks in the middle part of the p_e range. We also determined the optimal distribution of edge detectabilities p(p_e) for the vote histogram of the dense annotation (**Supplementary Fig. 4**). Again, we found the same general structure, with a strong peak near 1 and several peaks throughout the rest of the range (**Fig. 3h**, right).

Computing the consensus skeletons

We next used the annotation-decision model to find the consensus skeletons. We estimated (Eqn. 2) the edge detectability p_e (more precisely, its distribution) for each edge, given the agreeing and disagreeing votes. We made the assumption that true connectivity results in above-chance edge detectability ($p_e > 0.5$). This implies that the annotation decisions will converge towards the ground truth as the number of redundant annotations increases. When training the annotators, we encouraged this by providing training examples rich in difficult locations.

This assumption about the relationship between the detectability of an edge p_e and its actual connectedness is likely not to be entirely correct. The number of locations for which this assumption is incorrect is, however, likely to be small (the crossover region between the sketched curves in **Fig. 3k**, middle panel).

We, therefore, based our consensus rule for an edge on whether the estimated distribution of edge detectability given the agreeing and disagreeing votes cast for that edge, $p(p_e|(T,N))$, indicated that the edge at that location was more likely to be detected than not. (**Fig. 4a**, for details see Methods, Eqn. 3). By evaluating this rule for all possible combinations of agreeing and disagreeing votes, we obtained the optimal decision boundary in the vote histogram between "eliminate edge" and "keep edge" (white line in **Fig. 4b**, note that this optimal decision boundary is substantially below the majority rule, i.e. edges with less than majority agreement are typically accepted). Since the consensus rule depends on the distribution of edge detectabilities $p(p_e)$, the optimal boundary is generally different for different neurite datasets (**Fig. 4b** top vs. bottom).

Because edge elimination splits some skeletons (**Fig. 4c**), it is necessary to determine which skeleton pieces still belong together. Whenever annotators had started from a soma, we simply checked whether there was still a connection between the skeleton pieces and a seed region in the proximal dendrite (**Fig. 4c**). **Figure 4e** shows how for the 50-fold-annotated cell, the consensus skeleton now lacks a large number of (presumably) erroneous neurites. In other cases, multiple annotators were instructed to start at different seed points along the same neurite (**Supplementary Figs. 4,5**, see Methods). There, finding the consensus skeletons is substantially more complicated, but our model still yields reasonable consensus skeletons. Note that each consensus skeleton is actually a bundle of closely spaced skeleton pieces (**Fig. 4e**, right panel).

Annotator quality

So far we have assumed that the error rates of different annotators are similar. To determine how much error rates vary across annotators, we assessed for each annotator how close his/her skeletons ran to those of others by calculating 1) the average number of total votes for or against that annotator. This measure, when low, indicates that an annotator had followed many neurites in little agreement with the other annotators; and 2) his/her average ratio of agreeing to total votes (**Fig. 4d**). For the majority of annotators, the average ratio of agreeing to total votes was between 95% and 98% (**Fig. 4d**). The worst performing annotator (circle in **Fig. 4d**, black skeleton in **Fig. 4e**, left) generated a skeleton with more than 4 times the total path length, even entering additional cells. The best annotators, on the other hand, had as few as 2 disagreements with the 50-fold consensus skeleton.

The residual error rates of *RESCOPed* skeletons

To estimate how many errors one would still have to expect in the consensus skeletons, we computed the error probabilities for each of the decisions to eliminate or accept an edge. As described above, an edge is eliminated whenever the vote count for this edge indicated that it was *more* likely than not that the edge was incorrectly annotated. However, there remains an error probability that the edge was in fact correctly annotated and should have been accepted. To calculate the error probabilities for eliminated edges and accepted edges we integrated the distributions of edge detectability given the agreeing and disagreeing votes cast for that edge, $p(p_e|(T,N))$, for $p_e>0.5$, and $p_e<0.5$, respectively (**Fig. 4a**). Because the distribution of edge detectability given the votes $p(p_e|(T,N))$ becomes more sharply peaked as the total number of votes increases (**Fig. 5b**), the error rate for a given ratio of agreeing to total votes decreases.

As the number of annotators rises the accuracy of the consensus skeleton increases (**Fig. 5c**) initially steeply but then more slowly. The reason for this slowing is that, as the detectability of an edge approaches 0.5 the number of votes needed to achieve a given error rate diverges (edges with an edge detectability of exactly 0.5 are fundamentally undecidable). Therefore, near an edge detectability of p_e=0.5 the error for a large number of votes N is very sensitive to the shape of the probability distribution, p(p_e), and the error predictions for a large number of votes can scatter substantially for different neurites, or even different groups of annotators (s. **Supplementary Fig. 5d**).

We next compared this error-rate prediction with the actual accuracy of the consensus skeletons. We randomly selected from the 50 skeletons sets of 25, 10, 5, and 1 skeletons (n=6, 15, 20, 10, respectively) and computed the consensus skeleton for each set independently (**Fig. 5a**). Then we visually assessed the differences between all those consensus skeletons and the 50-fold consensus skeleton (which we took as reference). We found the average number of disagreements to be 1.0±0.4, 2.1±0.3, 7.2±0.9, and 15.5±3.5 (mean ± s.e.m.) for the 25-fold, 10-fold, 5-fold and single skeletons respectively, corresponding to mean distances between errors of 600.2 µm, 281.3 µm, 83.4 µm and 38.7 µm (**Fig 5c**, top panel).

So far we have considered the case where the entire length of neurites is multiply annotated. Since for most locations connectedness is easy to determine, increasing the overall redundancy is wasteful. We, therefore, explored focused re-annotation: repeatedly examine each edge until a given accuracy (which RESCOP provides) is reached rather than annotating each edge a fixed number of times. This should concentrate the annotators' effort onto difficult locations. In order to determine the redundancy-accuracy tradeoff for focused re-annotation we performed Monte-Carlo simulations and found that for focused re-annotation the accuracy should rise much more steeply, almost exponentially, with the average redundancy (**Fig. 5c**).

Variation of error rate with data quality

To test how the error rate depends on the staining method and on the data quality, we annotated a conventionally stained data set (K0563, s. Supplemental image stacks) at its original resolution (12×12×25 nm^3 voxels), with added noise (Gaussian, s.d.=20, original gray value range (101, 196), 3^{rd} and 97^{th} percentile, respectively), and at half the resolution (24×24×50 nm^3 voxels) (**Fig. 5d**). We found that error rates were actually slightly lower for the added-noise case, possibly due to increased attention, but that for the reduced-resolution data annotation reliability was substantially degraded (**Fig. 5c**, lower panel).

Dense reconstruction

To illustrate the feasibility of dense neuron reconstruction from SBEM data using the tools presented here, we selected all rod bipolar cells (RBCs, **Fig. 6**) from a SBEM data set that is in the process of being skeletonized (data set E2006, currently at 2 fold redundancy, Helmstaedter et al., in preparation). The E2006 data set covers a different block of tissue (sized 80 µm x 117 µm x 135 µm, s. Methods), came from a mouse rather than a rabbit retina, was imaged at a higher resolution, and stained more intensely. RBCs were initially identified on the basis of geometrical parameters using automatic clustering (Helmstaedter et al., in preparation). The selection was then refined by manually removing 23 of 137 cells because they were obviously cone bipolar cells (14 cells) or had an aberrant morphology, which indicated a substantial annotation error (9 cells), not yet eliminated due to the only 2-fold redundancy. The remaining 114 cells displayed the tiling patterns of axons and dendrites expected for rod bipolar cells (**Fig. 6c,d**). The annotation speed for these skeletons was 5.3 h per mm path length (the RBCs had an average neurite length of 368±103 µm, mean±s.d.). Using the model described above, we expect about 10 errors per cell for double annotation. To reduce the error rate to one per cell, a redundancy of 18 or 19 (and a redundancy of 4 on average for focused re-annotation) should be sufficient.

These numbers indicate that it is feasible to reconstruct all bipolar cells and all the dendrites or dendrite fragments of all ganglion cells with their somata in such a block of tissue using ~7500 work hours (s. Methods).

DISCUSSION

Dense vs. sparse reconstruction

Our data show that the dense reconstruction of neurites in SBEM volume electron microscopy data is feasible, but also that manual annotations contain errors, even when performed by experts. Many of these errors are caused by insufficient attention, particularly where neurites branch (**Figs. 2,3**). This problem does not occur when the labeling is sufficiently sparse but is prevalent for densely stained tissue, as is needed for any kind of connectomic analysis of neurite networks. Branching-type errors are likely to occur even for light-microscopic data as soon as the stained-neurite density is so large that the frequency of close encounters between neurites becomes substantial, as it does with even a moderate fraction of neurons stained. It is likely that annotation errors are widespread, but they are rarely acknowledged, let alone quantified. Annotation error rates are clearly related to the information content and quality of the staining (**Figs. 2,5**). For the study of local synaptic geometry, where a solid body of serial EM studies exists, a modest error rate will only rarely affect the conclusions. Error rates need, however, to be much lower for connectomic neuroanatomy, where a single missed branch point typically means thousands of lost or wrongly attributed synapses. Other errors are less costly; a missed spine neck would mean the loss of a few synapses at most. We have so far only quantified errors caused by incorrect neurite reconstruction. While the identification of synapses can be error-prone as well, one such error affects only one particular synapse, with a much less severe effect on the connectomic reconstruction error than the typical neurite continuity error has.

The few published reconstructions of entire neurites from EM data were performed by highly trained and dedicated experts, and extensively proof-read by the same or other experts[3, 11, 19-20]. For the *C. elegans* connectome a number of corrections were published recently[21], using the original image series. Some form of proof-reading clearly is necessary

during the connectomic reconstruction of neuronal networks[22-23]. Proofreading existing skeletons is, however, not only very tiresome (Helmstaedter et al., unpublished) but may well be less efficient than redundantly (multiply) annotating the same neurites followed by the detection of inconsistencies. Different from conventional proofreading, redundant annotation also allows the quantification of the annotation difficulty (**Fig. 5c**).

Mass annotation, distribution of skill and training levels

Finding the consensus of multiple annotations using RESCOP may reduce the error rate to a level sufficient for almost any application of connectomic circuit reconstruction. In addition, RESCOP estimates of the number of reconstruction errors remaining in the consensus skeleton, and points to likely locations for those errors, a prerequisite for focused re-annotation. Our analysis also shows that the optimal vote threshold (the decision boundary) can be substantially different from majority voting (**Fig. 4b**).

RESCOP allows the creation of connectomic reconstructions with a known accuracy while using annotators that have no prior neurobiological knowledge and are only slightly trained. Even if the error rate is rather high for individual annotators, it should be possible to reduce the error rate in the consensus skeleton substantially, with only a moderate increase in the average redundancy if focused re-annotation is used. Most of the effort could then be concentrated onto difficult locations (p_e near 0.5) which require a higher redundancy to reach a given reliability. In our data, difficult locations appear to be rare, as the prevalence of vote ratios near one shows (**Fig. 3**). The low density of difficult locations also means that ambiguous vote ratios (T/N near 0.5) are rare and the fits for p(p_e) are not very well constrained in the region around p_e=0.5, making estimates of error rates for large N somewhat uncertain (**Supplementary Fig. 5**). Ultimately, the error rate will be affected by the assumption we made that an infinite number of annotators will converge to the

correct decision. This limits the validity of the accuracy predictions (**Fig. 5c**) for very large N. We expect that the availability of improved staining and imaging methods will further reduce the frequency of locations at which the data biases even the expert towards the wrong conclusion.

One advantage of using weakly trained annotators is that the reliability increase can be achieved at a lower cost than with expert proof readers, who might still make attention-related errors at an unacceptable rate (**Fig. 2**); Also, requiring PhD-students or post-docs to do several thousand hours of annotating is hardly a good use of their talents. Finally, untrained personnel can in many academic settings be recruited quickly and on a temporary basis. The ability of RESCOP to automatically direct annotator effort and assess annotator quality makes it well suited for web-based crowd sourcing. All this makes it practical to scale the annotation capacity up, limited only by the available budget. RESCOP thus removes a major obstacle to high-throughput circuit reconstruction. This is demonstrated by our reconstruction of bipolar cells (**Fig. 6**), which took 5.3 h per mm of skeleton length, roughly 60 times faster than volume labeling.

Evaluation of automated reconstruction algorithms

Computer algorithms, especially those using machine learning[24-26] can help with the reconstruction of neural circuits. In the long run, such tools may well replace or at least greatly reduce the need for manual annotation. However, automatic methods need to be evaluated by comparing them to a reliable ground truth. The consensus among manual annotations can serve as such a ground truth, in particular when, as is the case for RESCOP, the error rate is known. Such an estimation of annotator errors is, incidentally, not available for other major benchmark datasets in machine learning (e.g., Berkeley Segmentation Dataset[27]). In medical imaging (MRI, CT), expert annotation (by trained radiologists) is the "gold" standard, but those expert annotations frequently differ

substantially[28]. Therefore, algorithms to estimate optimal annotations have recently received increasing attention (e.g., STAPLES[29]). Often, majority voting is found to be close to optimal[30]. The approach presented here is a "decision theoretic" one, which means finding the optimal decision criterion given a model of the belief formation (or decision) process[31].

Combining skeletons with automated methods

One major shortcoming of skeleton annotations is that they do not produce a complete volume representation, which is especially important for detecting contacts between neurites, a prerequisite for synapses. This problem can, however, be solved (Helmstaedter et al., unpublished) by combining high-accuracy long-range manual annotation, as reported here, with locally accurate but globally error-prone automated volume reconstructions[24-26]. Such hybrid techniques should reduce the manual effort to create full volume representations by as much as two orders of magnitude and will thus enable the connectomic reconstruction of much larger volumes than previously possible.

Acknowledgements:

The authors would like to thank Björn Andres, Fred Hamprecht, Ullrich Köthe, Viren Jain, Sebastian Seung, and Srinivas Turaga for many fruitful discussions and comments on the manuscript, Johann Bollmann for helpful comments on the manuscript, Chris Roome for IT support, Jörgen Kornfeld and Fabian Svara for programming KNOSSOS, and Janina Hanne, Hannah Jakobi, and Heiko Wissler for help with annotator training. We thank N Abazova, E Abs, A Antunes, P Bastians, J Bauer, M Beez, M Beining, S Bender, S Best, L Brosi, M Bucher, E Buckler, J Buhmann, C Burkhardt, F Drawitsch, L Ehm, S Ehm, C Fianke, R Foltin, S Freiss, M Funk, A Gebhardt, M Gruen, K Haase, J Hammerich, J Hanne, B Hauber, M Hensen, L Hofmann, P Hofmann, M Hülser, F Isensee, H Jakobi, M Jonczyk, A Joschko, A Juenger, S Kaspar, K Kessler, A Khan, M Kiapes, A Klein, C Klein, S Laiouar, T Lang, L Lebelt, H Lesch, C Lieven, D Luft, E Moeller, A Muellner, M Mueller, D Ollech, A Oppold, T Otolski, S Oumohand, S Pfarr, M Pohrath, A Poos, S Putzke, J Reinhardt, A Rommerskirchen, M Roth, J Sambel, K Schramm, C Sellmann, J Sieber, I Sonntag, M Stahlberg, T Stratmann, J Trendel, F Trogisch, M Uhrig, A Vogel, J Volz, C Weber, P Weber, K Weiss, L Weisshaar, E Wiegand, T Wiegand , M Wiese, R Wiggers, C Willburger, and A Zegarra for neurite skeletonizations.

Author contributions:

M.H. and W.D. designed the study and devised the analysis algorithms; K.L.B. performed the SBEM experiments; M.H., K.L.B., and W.D. specified the KNOSSOS software; M.H. analyzed the data; M.H., K.L.B., and W.D. wrote the paper.

FIGURE LEGENDS

Figure 1 Comparison of volume and skeleton annotation. Examples of volume labeling (**a**) and skeletonization (**b**) for the same two neurite fragment; cell-surface labeled data (dataset E1088, s. Methods). (**c**) Sketch of a neurite skeleton. (**d**) Rate of time consumption for volume labeling[10] and for skeleton annotation (data from this study; annotated using KNOSSOS, s. **Supplementary Movie**), for both cell-surface labeled data (black) and the conventionally stained dataset (K0563, gray, s. **Fig. 5d**). Error bars: range (volume labeling), s.d. (skeletoniziation). (**e**). Outline of the Redundant Skeleton COnsensus Procedure (RESCOP) as described in this article. Scale bar, 250 nm (a,b).

Figure 2 Skeletonization by expert annotators. (**a**) Two complete skeletons of the same amacrine cell annotated independently by MH and KB, starting at the soma. (**b**) Same skeletons shown looking onto the plane of the retina. Green indicates agreement between the annotators and black disagreement (numbers indicate disagreement locations). For stacks of original data surrounding the disagreement locations see **Supplementary Material**. INL, inner nuclear layer, IPL, inner plexiform layer, GCL, ganglion cell layer. Scale bars, 5 µm.

Figure 3 RESCOP, step 1: skeleton-to-skeleton agreement measurement. (**a**) Overlay of 7 independent skeletons of the same neurite (bipolar cell axon) annotated by weakly trained non-experts, all starting at the soma (red cross). (**b**–**e**) The procedure used to measure the agreement between multiple annotators, shown schematically for one skeleton edge (dashed line) in skeleton A. (**f**) Histograms of edge votes for the 50-fold annotation of one cell (left panel) and the dense skeletonization of 98 neurites (right panel). Below: vote count vs. total number of votes (log scale). Histograms were corrected for multiple counting of the same location, s. Methods. (**g**) Predicted vote histograms for the single cell (left) and for dense skeletons (right), using the distribution of edge detectabilities $p_{fit}(p_e)$ (**h**) that best predicted the respective histograms in f. (**i,k**) Schematic illustration of how the truth (top panels) is converted to detection probability (middle panels). Bottom panels: the probabilities for different T (number of pro votes) for one edge (i, binomial distribution for p_e=0.7 and N=10 annotators); and for all edges combined (k, schematic).

Figure 4 RESCOP, steps 2 and 3: edge elimination and skeleton recombination. (**a**) Probability that the edge detectability p_e has a certain value, given different edge votes, without prior knowledge (blue) and for the fitted distribution of edge detectabilities $p_{fit}(p_e)$ (red). Whether an edge is kept or eliminated depends on whether the integral of $p(p_e|T,N)$ for $p_e > 0.5$ (green shading) is larger or smaller than that for $p_e < 0.5$ (red shading). In this example edges with one agreeing vote (T=1) and 4 total votes (N=4) would be eliminated, those with T=2/4 to 4/4 would be kept. (**b**) Decision error, $p_{err}(T,N)$, with optimal (stepped line) and majority vote (dashed straight line) decision boundaries for the single-cell data (top) and the dense skeletonization data (bottom). (**c**) Elimination procedure illustrated at a branch point. Red, eliminated edges. Green, discarded skeleton pieces (**d**) Variation of annotator performance as reflected in the average total number of votes per edge and the average ratio of agreeing to total votes for each annotator. Circle: worst-performing annotator who skeletonized the black skeleton in (e). (**e**) 50 skeletons of one amacrine cell before (left) and after (right) edge validation and consensus computation. Scale bar, 5 µm.

Figure 5 RESCOP, step 4: estimating the error-rate of RESCOP-ed skeletons. **(a)** Stereo view of two superimposed sets (red, blue) of 5-fold consensus skeletons. Black asterisks indicate disagreements. Total neurite path length: 600 µm. **(b)** Estimated detectability distribution for one edge for a fixed ratio of agreeing to total votes (*T/N)* of 0.33, but different numbers of total votes *(N)*. Probabilities are given that the edge was erroneously kept. **(c)** Top panel: mean path length between errors as a function of the number of annotators. Solid lines: estimates using Eqn. 11, for the dense neurites (red) and for the single cell (green)**,** crosses: errors detected by visual comparison with the 50 fold consensus skeleton for the consensus of 1, 5 (includes a), 10, and 25 skeletons (error bars, s.e.m.). Dashed lines: the average redundancy as a function of the target error rate for focused re-annotation (Monte-Carlo simulations). Bottom panel: Same analysis for a conventionally stained dataset annotated using the original data (blue), with added noise (magenta), and at half the resolution (cyan). **(d)** Examples from the original and degraded datasets. Scale bar, 250 nm.

Figure 6 Doubly annotated skeletons of 114 putative rod bipolar cells in a block of mouse retina. (**a**) View onto the block face. (**b**) The two skeletons for a single rod bipolar cell. (**c**, **d**) view onto the plane of the retina confined (as indicated in a) to the dendrites (c) and axons (d) of the bipolar cells, respectively. Cells are colored randomly in c,d. Note how, for the most part, the dendrites and axons tile the space. Scale bars, 10 μm.

METHODS

SBEM

The retinas (from a 6-week old rabbit for E1088, used in **Figs. 1–5** and **Supplementary Figs. 1,4**, from a P30 C57BL/6 mouse for E2006, used for the data in **Fig. 6** and **Supplementary Fig. 2**, and from a P30 C57BL/6 mouse for K0563 were prepared for E1088 and E2006 to selectively enhance cell outlines by using HRP-mediated precipitation of DAB (for preparation details[16]) stained with osmium alone (E1088) or in conjunction with lead citrate (E2006). For K0563 a more conventional stain was used (same dataset as in[16]). All procedures were approved by the local animal care committee and were in accordance with the law of animal experimentation issued by the German Federal Government.

The embedded tissue was trimmed to a block face of approximately 200 µm x 300 µm in size, and imaged in a scanning electron microscope with a field-emission cathode (QuantaFEG 200, FEI Company, Eindhoven, the Netherlands) and a custom-designed back-scattered electron detector based on a special silicon diode (AXUV, International Radiation Detectors, Torrance CA) combined with a custom-built current amplifier. The incident electron beam had an energy of 3.6 keV and a current of ~100 pA for E1088, an energy of 3.0 keV and a current of ~100 pA for E2006, and an energy of 2.0 keV and a current of ~100 pA for K0563. At a pixel dwell time of 8 µs and a pixel size of 22 nm x 22 nm (E1088), 6 µs and 16.5 nm x 16.5 nm (E2006), and 5 µs and 12 nm x 12 nm (K0563), this corresponds to doses of about 10 (E1088), 14 (E2006), and 22 (K0563) electrons per nm^2, not accounting for skirting due to low vacuum operation. The chamber was kept at a pressure of 75 Pa of water vapor (E1088) or 130 Pa of hydrogen (E2006) to prevent charging. K0563 was conducting enough to be imaged in high vacuum. The electron microscope was equipped with a custom-made microtome similar to the one described in

[12], which allows the repeated removal of the block surface at ~30 nm (E1088) cutting thickness (~25 nm for E2006 and K0563). 1999 (for E1088), 3200 (for E2006), and 5765 (K0563) consecutive slices were imaged, resulting in data volumes of 2048×1768×1999 voxels (E1088), 8192×7072×3200 voxels (4×4 mosaic of 2048×1768 images, E2006), and 4096 × 5304 × 5760 voxels (2 × 3 mosaic of 2048 × 1768 images, k0563), corresponding to volumes of 45×39×60 μm^3, 135×117×80 μm^3, and 50×65×145 μm^3. For E1088 the imaged region spanned the inner plexiform layer of the retina and included parts of the inner nuclear and of the ganglion cell layers. E2006 spanned the retina from the ganglion cell layer to the cell bodies of photoreceptors. K0563 spanned the inner plexiform layer of the retina and included the ganglion cell layer and part of the inner nuclear layer. Consecutive slices were aligned off-line to sub-pixel precision by Fourier shift-based interpolation, using cross correlation-derived shift vectors.

Reconstruction software

Neurite skeletons were annotated using KNOSSOS (written in C by Jörgen Kornfeld and Fabian Svara according to specifications by the authors). KNOSSOS (s. Supplementary Movie) will be available for download after publication.

Skeletonization

Data were annotated using KNOSSOS and skeletons saved in an .xml format (called .nml), very similar to the NeuroML format[32]). Each file contains a list of the skeleton nodes, for each node a number of parameters (including index, coordinates, radius, viewport used for node placement, timestamp) is given, a list of the edges between nodes, and a list of nodes tagged as branch points (an example file is provided in the Supplement). Annotators were instructed as follows: (1) start at a given seed point (typically inside the soma of a neuron, for randomly dense seeding strategies, see below and Suppl. Material); (2) follow the neurite from that location, note that the neurite generally continues in two (if the seed point is in an axon or dendrite) or more (if the seed point is in a soma with more

than 2 primary neurites) directions; (3) while annotating, focus on the viewport that is most orthogonal to the current neurite axis (more recent versions (after v3.0435) of KNOSSOS determine the appropriate viewport, using the vector between the two most recently placed nodes, and highlight it); (4) accuracy is more important than speed; (5) place a node approximately every 7-10 planes (corresponding to ~200-300 nm edge length for SBEM data); (6) generously place branch-point flags, in order not to miss branches. Annotators were trained on at least 3 neurons (typically 10-40 hours of training). Their training results were compared to annotations of the same neurons by experts, and disagreements were inspected and discussed. Annotators were only allowed to continue with novel tasks when the training performance was sufficient as judged by the trainer.

Speed measurement

To measure the speed of skeletonization we initially asked annotators to report the time spent annotating. This yielded 5-10h per mm of path length annotation time. Then we included a time stamps feature in KNOSSOS that records the time when each skeleton node is placed (**Supplementary Fig. 2a**). To determine the effective annotation rate we summed up the inter-node time intervals, excluded intervals longer than 7 minutes to account for breaks taken by the annotators. This assumes that no single location takes that long to contemplate (**Supplementary Fig. 2b,c**).

Edge validation algorithm

To test each edge in a given set of skeletons $\{ S_\alpha, S_\beta, S_\gamma, .. \}$, that were created by multiple annotators (α, β, γ, ..) starting at the same seed point (for different re-seeding strategies, s. below), we used the following procedure. To test, say, edge $E_{\alpha ij}$, (which connects nodes N_i and N_j in skeleton S_α), S_α was first pruned beyond a sphere of radius r_p around the center of the edge $E_{\alpha ij}$, yielding two skeleton pieces $S_{\alpha i1}$, $S_{\alpha i2}$ starting at the ends of $E_{\alpha ij}$ ($N_{i\alpha}$ and $N_{j\alpha}$, respectively, Piece 1 and 2 in **Fig. 3b**). The cutoff radius r_p was set to ensure that at least one further edge was included at each end of the tested edge:

$$r_p = \max\left(|E_{ki\alpha}|/2 + \max\left(\min_k(|E_{ki\alpha}|), \min_k\left(|E_{kj\alpha}|\right)\right), 625nm\right), \quad \textbf{(1)}$$

where $\min_k(|E_{kj\alpha}|)$ is the length of the shortest of the edges connected to node $N_{j\alpha}$ (for the 50-fold single cell voting, r_p was on average 28 voxels, i.e. ~700 nm). Next, one of the other skeletons, S_β, was taken and the root-mean-squared node-to-edge distances were calculated between each of the skeleton pieces $S_{\alpha i1}$, $S_{\alpha i2}$, and S_β using all nodes of $S_{\alpha i1}$ and $S_{\alpha i2}$. When both root-mean-squared distances were below a set threshold θ = 625 nm this was a vote in favor of the tested edge (the agreeing-vote count $T_{\alpha ij}$ and the total vote count $N_{\alpha ij}$ for edge $E_{\alpha ij}$ were both raised by one); if only one but not the other distance was below the threshold this counted against $E_{\alpha ij}$ (only the total vote count $N_{\alpha ij}$ for edge $E_{\alpha ij}$ was raised by one). For both distances above the threshold no vote was counted because this indicated that S_β was not near the tested edge. θ was on the order of the typical neurite radius, which, however, varies widely; both θ and r_p were selected so as to minimize the difference between the 50-fold-consensus skeleton and sets of 10-fold-consensus skeletons. If the edge was within 3 nodes of a neurite ending we used $\theta_{end}=2*r_p$ as the threshold for agreement to account for the variability in the placement of terminal nodes. This procedure was repeated for all remaining skeletons S_γ, S_δ, ..., and T and N were finally both raised by one to account for the tested edge itself (seen as agreeing with itself). While the reliability of consensus skeletons is likely to be lower near endings, errors near endings are also less consequential since the number of misallocated nodes is small.

Finding the consensus skeleton

After validating all edges, the consensus skeleton was computed. Finding the consensus skeletons means eliminating those edges that are more likely to be wrong than correct. In order to decide whether to eliminate or keep an edge, given a vote (T,N), we calculated the conditional probability distribution of the hidden parameter p_e (toward which T/N would converge for an infinite number of annotators):

$$\mathrm{p}(p_e \mid T, N) = \frac{\mathrm{p}(T \mid N, p_e)\,\mathrm{p}(p_e)}{\mathrm{p}(T \mid N)}. \quad \textbf{(2)}$$

Because we assumed that the annotators have no additional bias, an edge should be eliminated if and only if

$$\int_0^{0.5} \mathrm{p}(p_e \mid T, N)\,dp_e > \int_{0.5}^{1} \mathrm{p}(p_e \mid T, N)\,dp_e. \quad \textbf{(3)}$$

For independent annotators the model for the likelihood is the binomial distribution,

$$\mathrm{p}(T \mid N, p_e) = \binom{N}{T} p_e^T (1-p_e)^{N-T}. \quad \textbf{(4)}$$

To determine the most likely p*(p_e)*, we computed the predicted vote histograms, $hist_{pred}$, while varying p*(p_e),* and compared $hist_{pred}$ to the measured vote histogram, $hist_{meas}$, in the following way. First we needed to correct for the fact that if, at one given location, *T* of *N* skeletons agreed, there will be a vote entry at (*T,N*) in the histogram for *each* of the *T* skeletons. We, therefore, divided the vote counts by *T* ($hist^*_{meas}(T,N) = hist_{meas}(T,N)/T$). Since we cannot measure edges with *T*=0, the predicted vote distribution was normalized for *T*=1..*N*:

$$hist_{pred}(T \mid N) = \mathrm{p}\left(T \mid N, \mathrm{p}_{fit}\right) \frac{\sum_{T=1}^{N} hist^*_{meas}(T \mid N)}{\sum_{T=1}^{N} \mathrm{p}\left(T \mid N, \mathrm{p}_{fit}\right)}, \quad (5)$$

whereby $\mathrm{p}\left(T \mid N, \mathrm{p}_{fit}\right) = \int_0^1 \mathrm{p}\left(T \mid N, p_e\right) \mathrm{p}_{fit}\left(p_e\right) dp_e$ is the probability that an edge that was sampled *N* times has exactly *T* pro votes.

We then assumed $\mathrm{p}_{fit}(p_e)$ to be a function that is piecewise linear between the points ρ_i= **f**(*i*/80), with *i* running from 0 to 80 and **f**(*x*)=$2x^2$ for *x*<0.5 and **f**(*x*)=1-2$(1-x)^2$ for *x*>=0.5.

This ensures that $p_{fit}(p_e)$ is more finely sampled near 0 and also near 1, where the bulk of the probability mass is expected. We can write $p_{fit}(p_e)$ as a sum over triangle-shaped basis functions $\mathbf{g}_i$ with peaks at the points ρ_i and weights w_i

$$p_{fit}(p_e \mid w_0..w_{80}) = \sum_{i=0..80} w_i g_i(p_e), \quad \textbf{(6)}$$

leading to a vote prediction of

$$hist_{pred}(T,N) = \sum_i (w_i c_i) \sum_{T=1..N} hist^*_{meas}(T,N) \Big/ \left(1 - \sum_i (w_i c_{i,T=0,N})\right), \quad \textbf{(7)}$$

whereby $c_{i,T,N} = \int_0^1 p(T \mid N, p_e) g_i(p_e) dp_e$. We varied all w_i to maximize the probability,

$$\prod \frac{e^{-\lambda}\lambda^k}{k!} = \prod_{T>0,N} \frac{e^{-hist_{pred}(T,N|w_0..w_{80})} \left(hist_{pred}(T,N \mid w_0..w_{80})\right)^{hist^*_{meas}(T,N)}}{hist^*_{meas}(T,N)!}, \quad \textbf{(8)}$$

that a given prediction leads to the observed vote distribution, assuming a Poisson distribution for the individual votes (with λ the expected number of events, and k the actual number of events). This correctly weights even small histogram numbers (including zero). Fitting was implemented in both Matlab (Mathworks) and Mathematica (Wolfram Research), yielding identical results.

After edge elimination, we collected all skeleton nodes for all redundantly annotated skeletons that still were connected to a source seed area near the soma by a continuous path of edges (using connected components). This then constituted the *RESCOP* consensus skeleton. The remaining skeleton pieces were discarded. For methods to reuse the discarded skeleton pieces, especially for locally dense skeletonization, see below and Supplemental Material.

Accuracy of *RESCOP*-ed skeletons

The calculation made to decide whether or not to eliminate an edge can be extended to calculate the probability that the decision was wrong and that the RESCOPed consensus skeleton therefore contains an error at that point.

For a given (T,N) the probability that $p_e>0.5$ is

$$\mathrm{p}_{keep}(T,N) = \int_{p_e=0.5}^{1} \mathrm{p}(p_e \mid T,N)\, dp_e \,. \qquad \textbf{(9)}$$

If the edge is kept, the probability of having done so erroneously is $1-\mathrm{p}_{keep}(T,N)$

Conversely, if the edge is eliminated the error probability is $\mathrm{p}_{keep}(T,N)$. The decision rule (Eqn. 3) to keep an edge if and only if $p_{keep}(T,N)>0.5$ minimizes the error probability

$$\mathrm{p}_{err}(T,N) = \mathrm{Min}\left(\mathrm{p}_{keep}(T,N), \left(1-\mathrm{p}_{keep}(T,N)\right)\right), \qquad \textbf{(10)}$$

and is thus optimal. To get the error rate for a given N we now need to sum this over T weighted by the probability ($\int_0^1 \mathrm{p}(T \mid N, p_e)\,\mathrm{p}(p_e)\, dp_e$) of T occurring.

$$\mathrm{p}_{err}(N) = \sum_{T=0}^{N} \left(\mathrm{p}_{err}(T,N) \int_0^1 \mathrm{p}(T \mid N, p_e)\,\mathrm{p}(p_e)\, dp_e \right), \qquad \textbf{(11)}$$

This is now the probability that there is still an error after finding the consensus of N skeletons at a given location. The mean path length between errors is then $r_p/p_{err}(N)$ (r_p was used rather than the edge length, because our voting procedure creates a correlation between errors of neighboring edges, s. **Fig. 3b–e**).

Focused re-annotation

To estimate the average annotation redundancy for the case where each edge is re-annotated until a given accuracy goal is reached we ran a Monte Carlo simulation as follows. We picked a p_e using $\mathrm{p}(p_e)$ as the probability density, repeatedly tossed a coin biased with p_e, and incremented T and N accordingly each time, until $\mathrm{p}_{err}(T,N)$ fell below

the set accuracy goal or N_{max} were reached. The set accuracy goal was then corrected for the residual errors for those runs that reached N_{max} (N_{max} = 6000, e1088 single-cell data and k0563 data, **Fig. 5c**, **Supplementary Fig. 5d**), with the exception of the dense skeletonization data where the number of runs that reached N_{max} was small. (N_{max} = 200, **Fig. 5c**, **Supplementary Fig. 5b**).

Random skeleton re-seeding

For the nearly dense reconstruction of neurites in a limited region (used in **Fig. 3f,g**) we did not seed at the somata, since those were not contained in the region, but used a random seeding/iterated re-seeding strategy (**Supplementary Figs. 1,4**). Briefly, annotation was restricted to a sphere around a seed point but seed-point placement was iterated several times, each time using as new seeds the end points of the skeletons from the previous iteration. To take into account that an enforced ending near a tested edge should not count against that edge while a natural endpoint should, *RESCOP* was appropriately modified. Another modification placed the skeletons remaining after edge elimination into clusters based on the proximity of the skeleton pieces. We also accounted for the possibility that some of the randomly placed initial seed points were in the same neurite. For details s. Supplement.

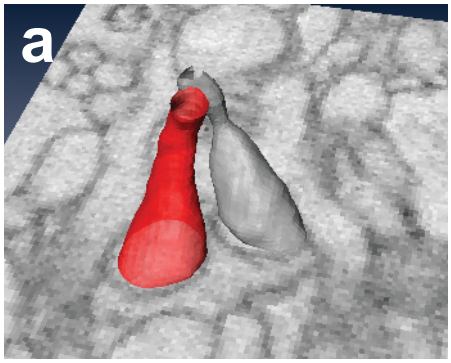
Reconstruction-cost estimation

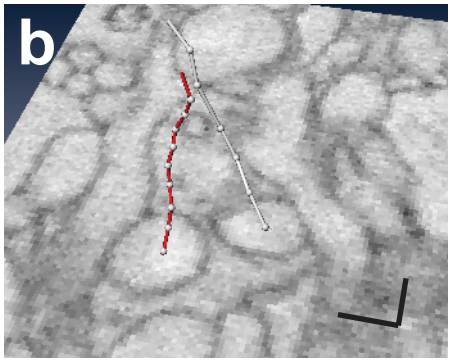
To calculate reconstruction costs we estimated that a block of mouse retina sized (120×80×130)µm^3 contains ~460 bipolar cells with ~0.3-0.8 mm path length each and ~40 ganglion cell somata with 1-2 mm dendritic path length each, which in most cases is only part of the dendrite. Annotating these at 6 h/mm with 4-fold redundancy will take 7500 work hours. In our setting each undergraduate student works on average 27 h per month. The reconstruction of all bipolar and ganglion cells at 4-fold redundancy would thus take 3 months with a team of 120 annotators.

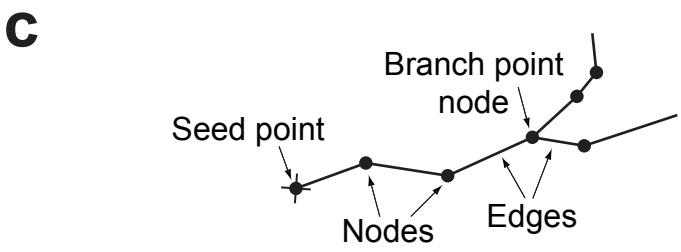
REFERENCES

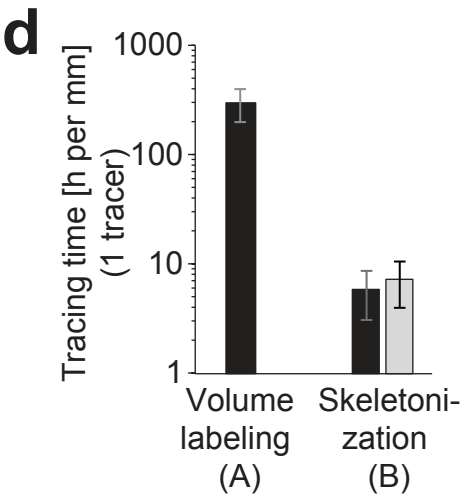
1. Ramón y Cajal, S. *Textura del Sistema Nervioso del hombre y de los vertebrados* (Moya, Madrid, 1899).
2. Golgi, C. Sulla struttura della sostanza grigia del cervelo. *Gazzetta Medica Italiana. Lombardia* **33**, 244-246 (1873).
3. Harris, K.M. & Stevens, J.K. Dendritic spines of CA 1 pyramidal cells in the rat hippocampus: serial electron microscopy with reference to their biophysical characteristics. *J Neurosci* **9**, 2982-2997 (1989).
4. Horikawa, K. & Armstrong, W.E. A versatile means of intracellular labeling: injection of biocytin and its detection with avidin conjugates. *J Neurosci Meth* **25**, 1-11 (1988).
5. Stretton, A.O. & Kravitz, E.A. Neuronal geometry: determination with a technique of intracellular dye injection. *Science* **162**, 132-134 (1968).
6. Wickersham, I.R*., et al.* Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. *Neuron* **53**, 639-647 (2007).
7. Livet, J*., et al.* Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450**, 56-62 (2007).
8. Sporns, O., Tononi, G. & Kotter, R. The human connectome: A structural description of the human brain. *PLoS Comput Biol* **1**, e42 (2005).
9. Lichtman, J.W., Livet, J. & Sanes, J.R. A technicolour approach to the connectome. *Nat Rev Neurosci* **9**, 417-422 (2008).
10. Helmstaedter, M., Briggman, K.L. & Denk, W. 3D structural imaging of the brain with photons and electrons. *Curr Opin Neurobiol* **18**, 633-641 (2008).
11. White, J.G., Southgate, E., Thomson, J.N. & Brenner, S. The Structure of the Nervous System of the Nematode Caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci* **314**, 1-340 (1986).
12. Denk, W. & Horstmann, H. Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. *PLoS Biol* **2**, e329 (2004).
13. Hayworth, K.J., Kasthuri, N., Schalek, R. & Lichtman, J.W. Automating the Collection of Ultrathin Serial Sections for Large Volume TEM Reconstructions. *Microsc Microanal* **12 (Supp2)**, 86-87 (2006).
14. Knott, G., Marchman, H., Wall, D. & Lich, B. Serial section scanning electron microscopy of adult brain tissue using focused ion beam milling. *J Neurosci* **28**, 2959-2964 (2008).
15. Briggman, K.L. & Denk, W. Towards neural circuit reconstruction with volume electron microscopy techniques. *Curr Opin Neurobiol* **16**, 562-570 (2006).
16. Briggman, K.L., Helmstaedter, M. & Denk, W. Wiring specificity in the direction-selectivity circuit of the retina. *Nature, in press* (2011).
17. Fiala, J.C. Reconstruct: a free editor for serial section microscopy. *J Microsc* **218**, 52-61 (2005).
18. Jeong, W*., et al.* Ssecrett and NeuroTrace: Interactive Visualization and Analysis Tools for Large-Scale Neuroscience Data Sets. *IEEE Computer Graphics and Applications* **30**, 58-70 (2010).

19. Trachtenberg, J.T.*, et al.* Long-term in vivo imaging of experience-dependent synaptic plasticity in adult cortex. *Nature* **420**, 788-794 (2002).
20. Stevens, J.K., McGuire, B.A. & Sterling, P. Toward a functional architecture of the retina: serial reconstruction of adjacent ganglion cells. *Science* **207**, 317-319 (1980).
21. Chen, B.L., Hall, D.H. & Chklovskii, D.B. Wiring optimization can relate neuronal structure and function. *Proc Natl Acad Sci U S A* **103**, 4723-4728 (2006).
22. Chklovskii, D.B., Vitaladevuni, S. & Scheffer, L.K. Semi-automated reconstruction of neural circuits using electron microscopy. *Curr Opin Neurobiol* **20**, 667-675 (2010).
23. Mishchenko, Y.*, et al.* Ultrastructural analysis of hippocampal neuropil from the connectomics perspective. *Neuron* **67**, 1009-1020 (2010).
24. Jain, V.*, et al.* Supervised Learning of Image Restoration with Convolutional Networks. *ICCV* (2007).
25. Andres, B., Köthe, U., Helmstaedter, M., Denk, W. & Hamprecht, F. Segmentation of SBFSEM Volume Data of Neural Tissue by Hierarchical Classification. in *Pattern Recognition* 142-152 (2008).
26. Turaga, S.C.*, et al.* Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural Comput* **22**, 511-538 (2010).
27. Martin, D., Fowlkes, C., Tal, D. & Malik, J. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. *ICCV* **2**, 416-423 (2001).
28. Warfield, S.K., Zou, K.H. & Wells, W.M. Validation of image segmentation by estimating rater bias and variance. *Philos Transact A Math Phys Eng Sci* **366**, 2361-2375 (2008).
29. Warfield, S.K., Zou, K.H. & Wells, W.M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* **23**, 903-921 (2004).
30. Wang, W.*, et al.* A classifier ensemble based on performance level estimation. *ISBI*, 342-345 (2009).
31. Kording, K. Decision theory: what "should" the nervous system do? *Science* **318**, 606-610 (2007).
32. Crook, S., Gleeson, P., Howell, F., Svitak, J. & Silver, R.A. MorphML: level 1 of the NeuroML standards for neuronal morphology data and model specification. *Neuroinformatics* **5**, 96-104 (2007).









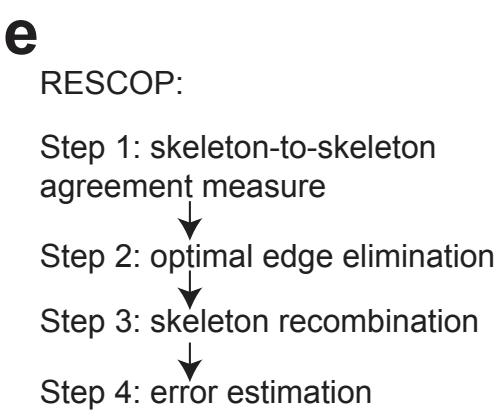


Figure 1
Helmstaedter et al.

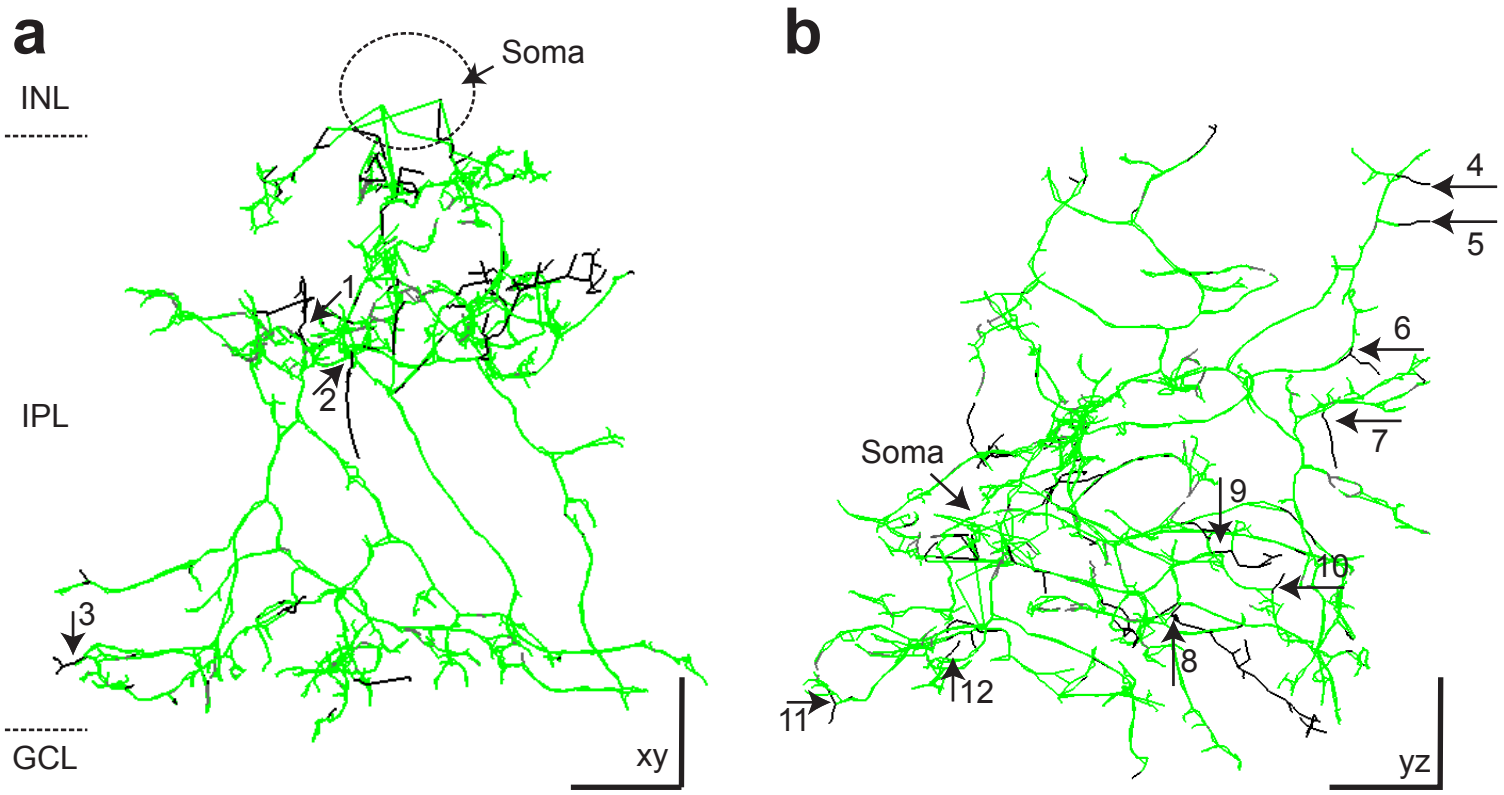


Figure 2
Helmstaedter et al.

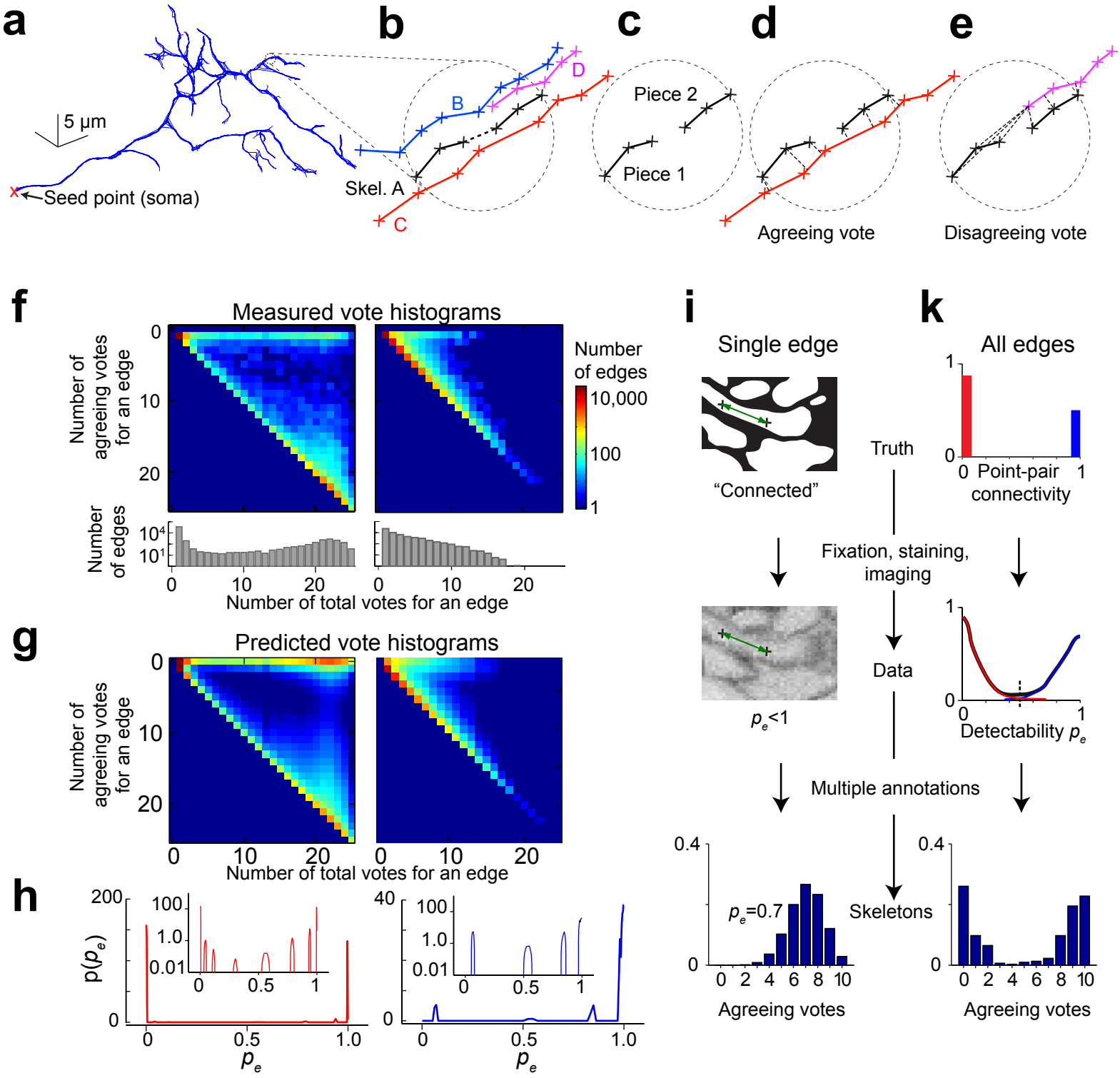


Figure 3
Helmstaedter et al.

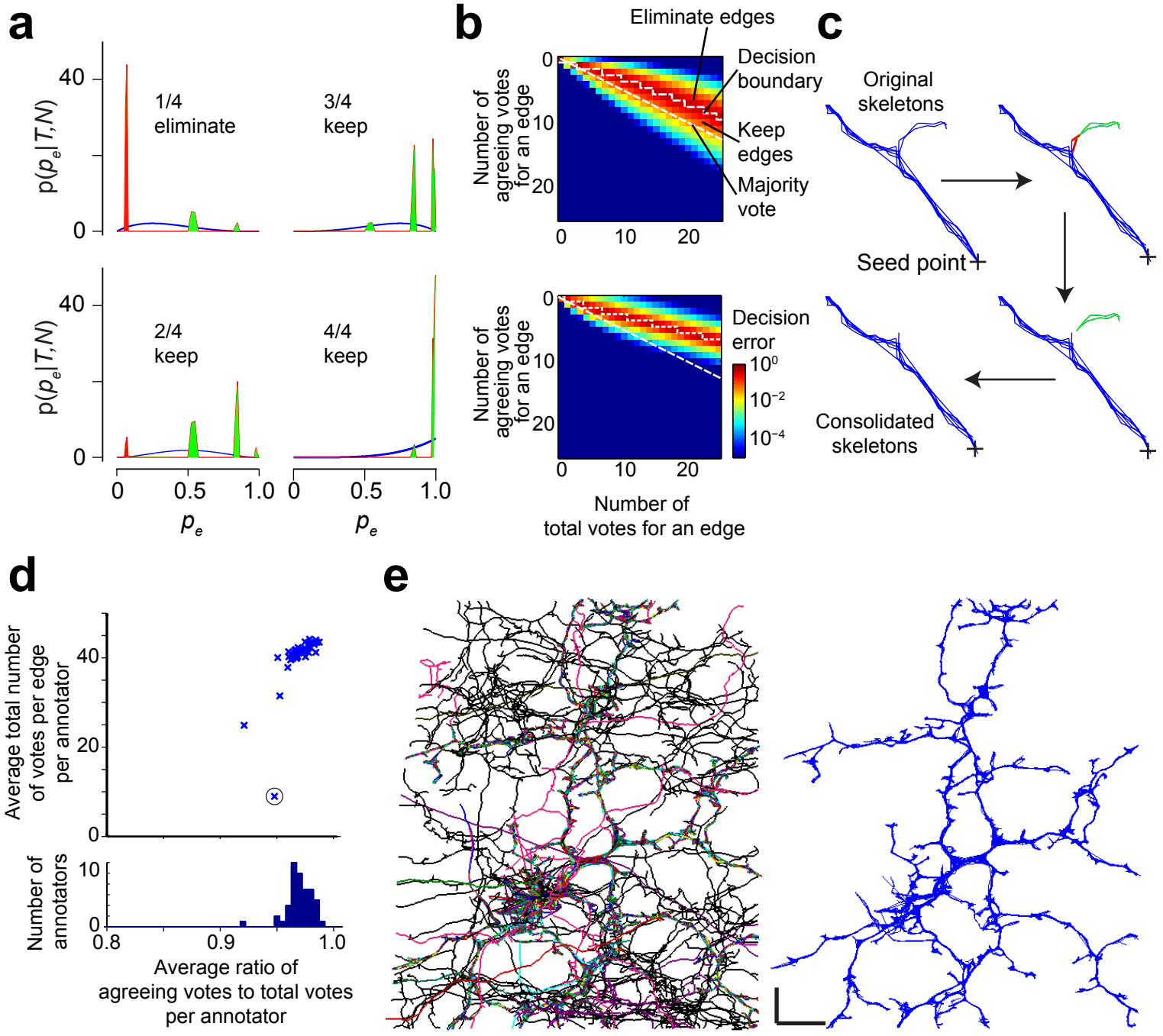


Figure 4
Helmstaedter et al.

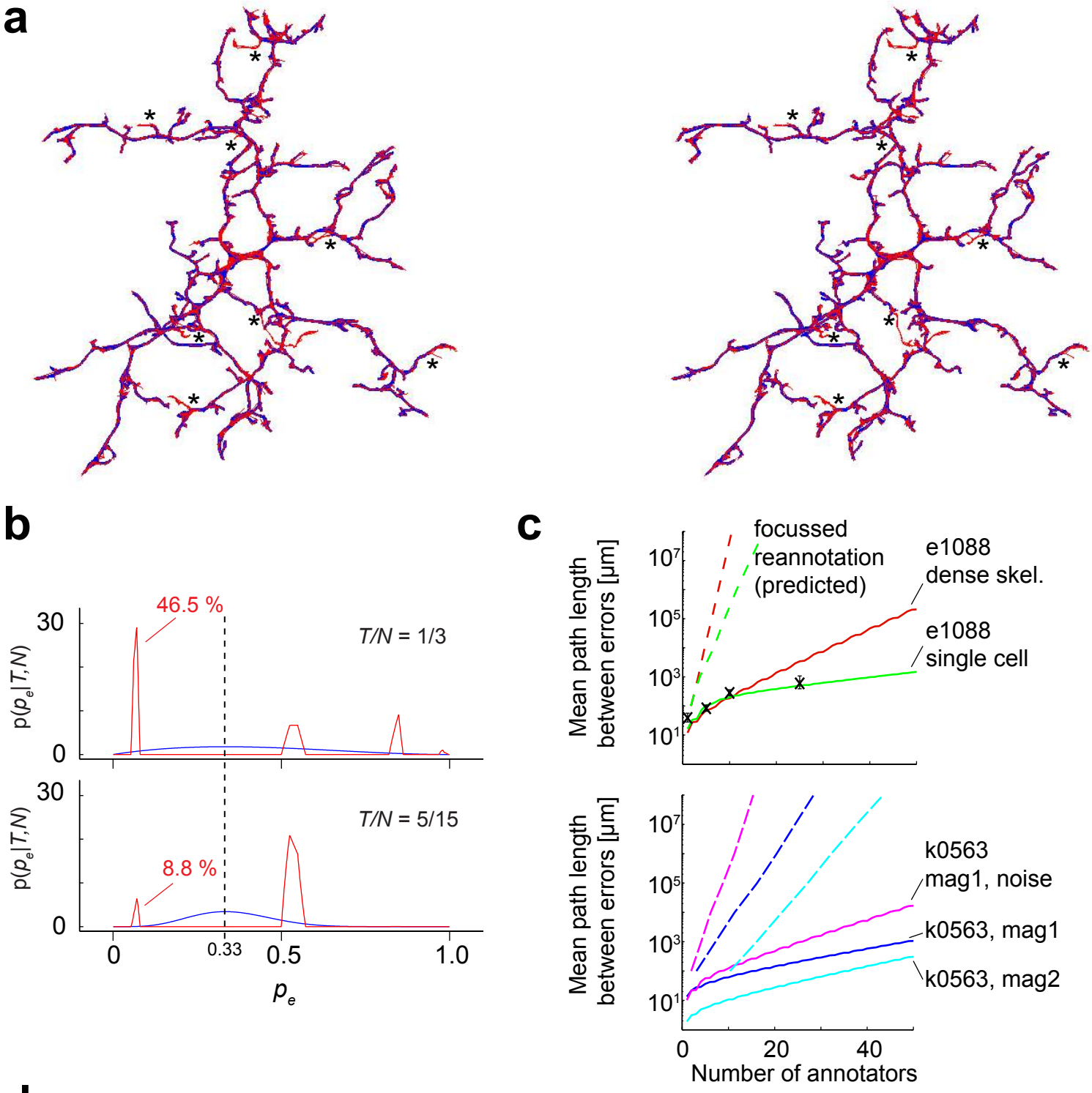


Figure 5
Helmstaedter et al.

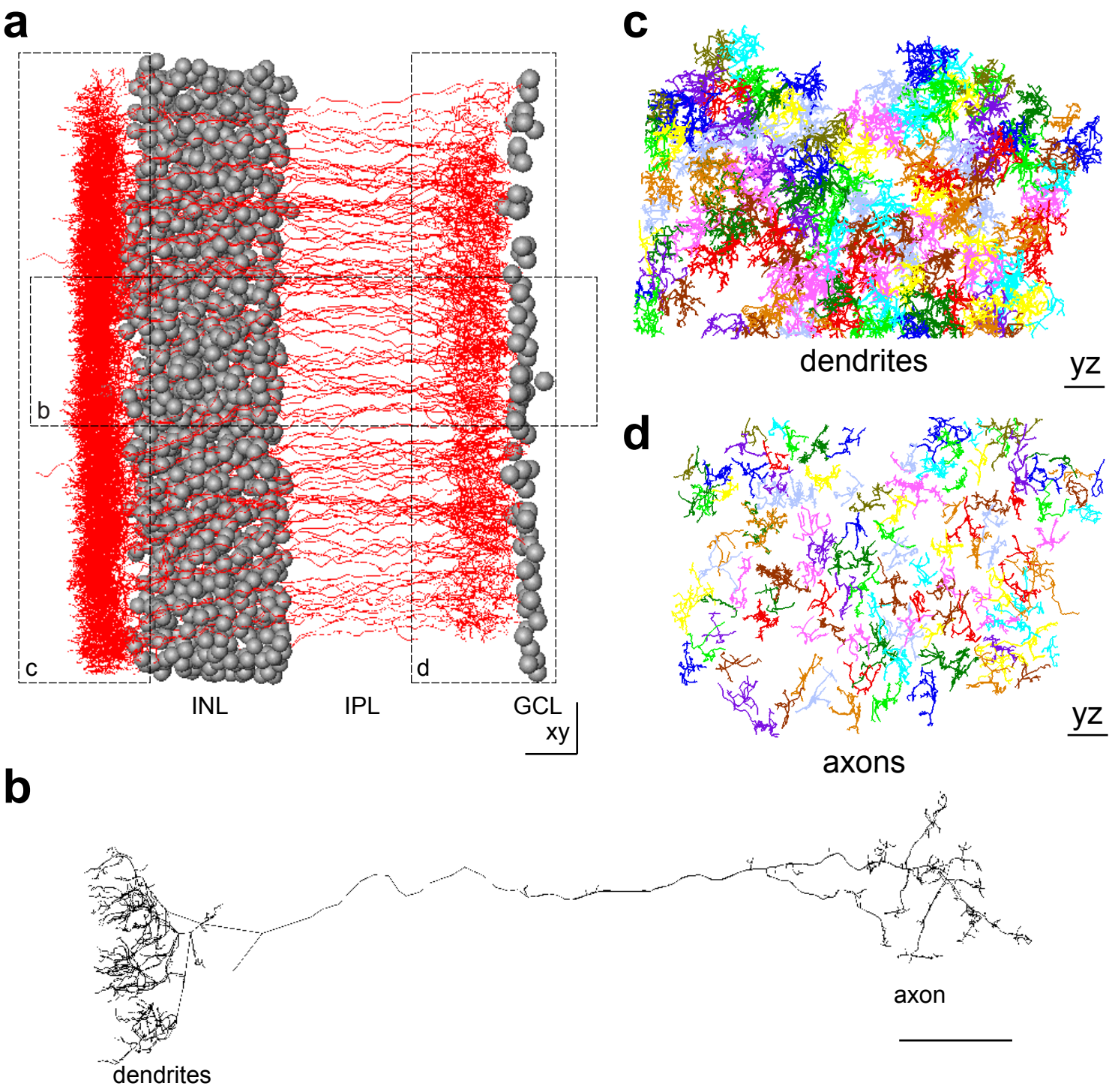


Figure 6
Helmstaedter et al.