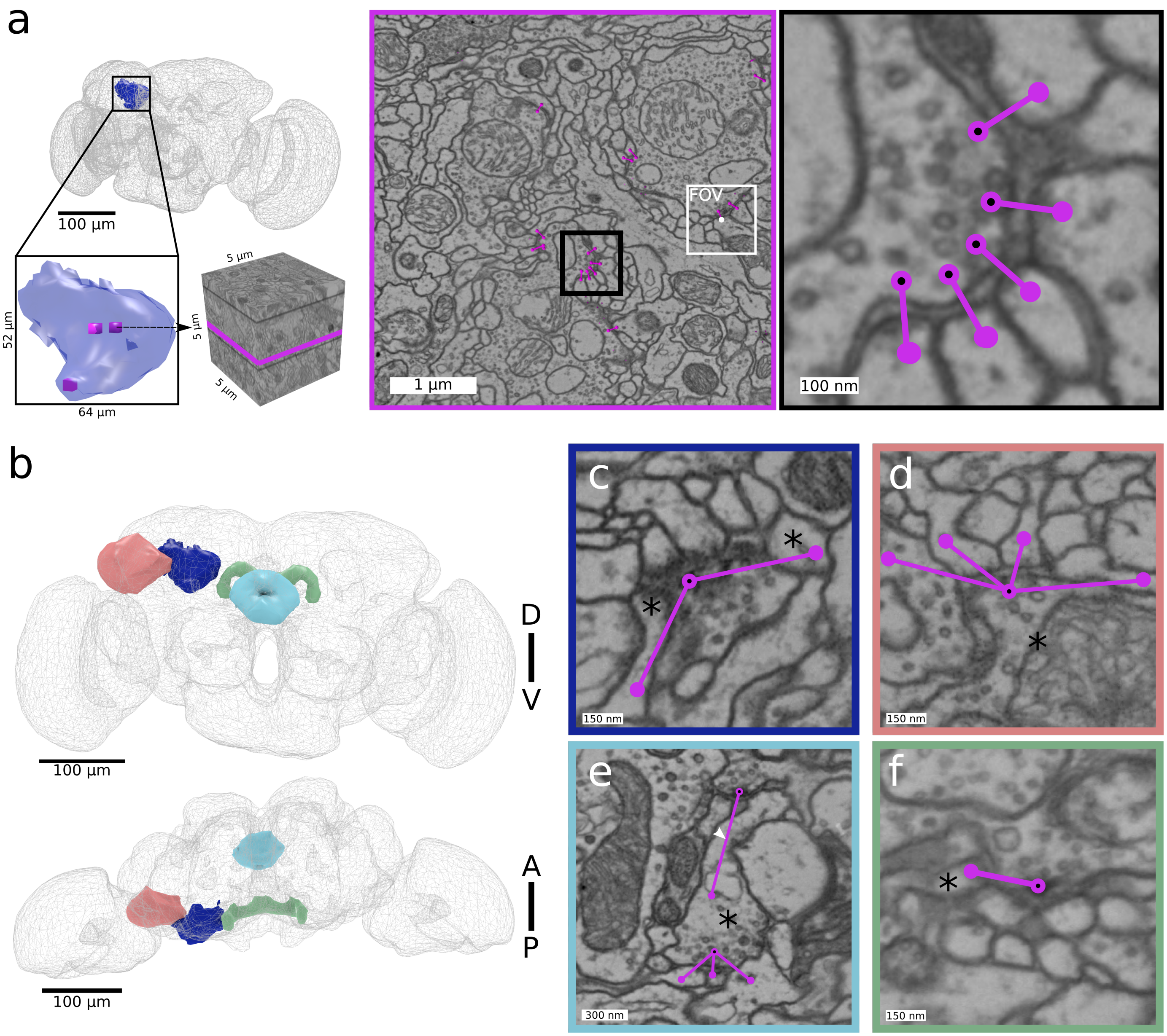
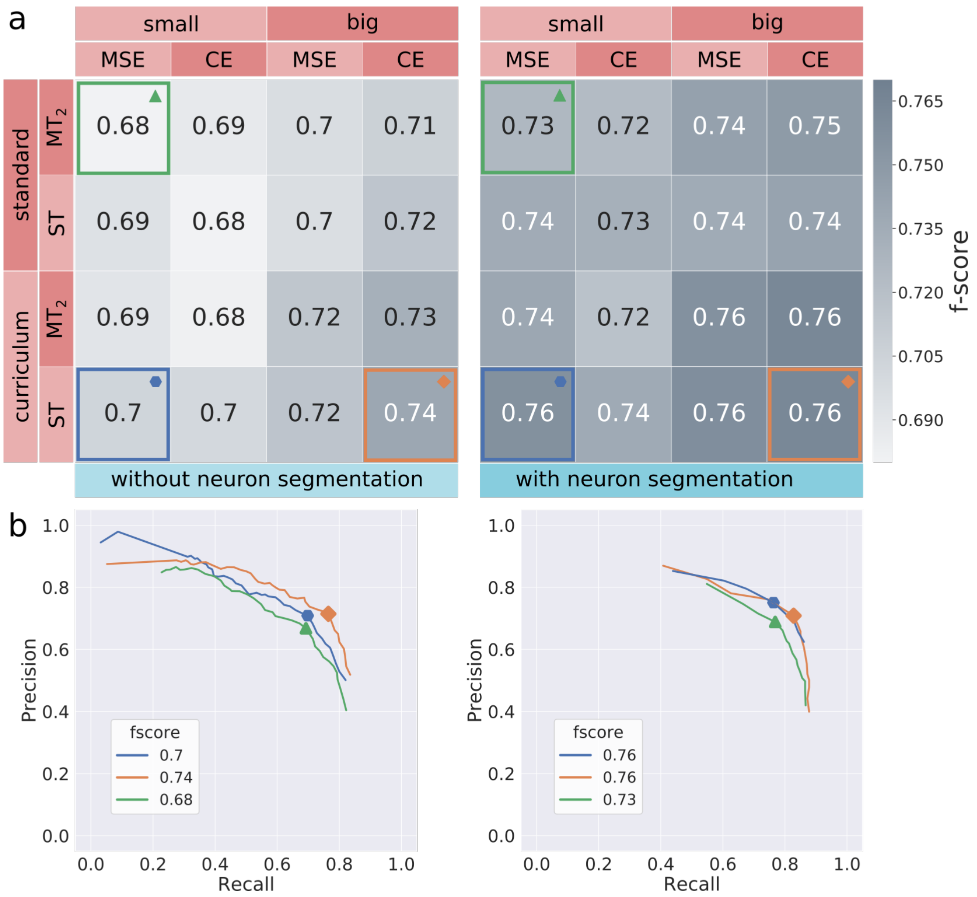
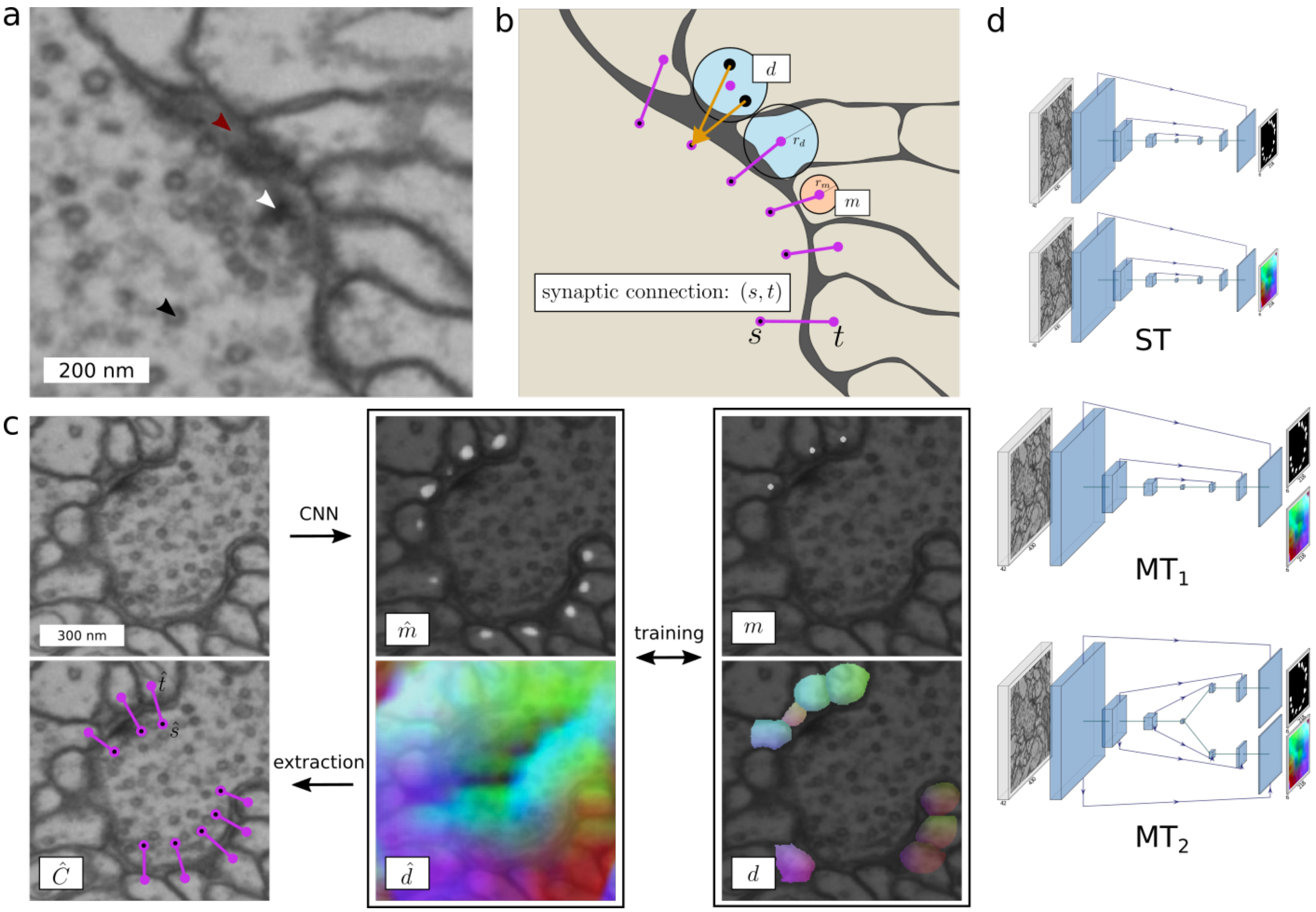


Supplementary Figure 1: Synapse ground-truth datasets in FAFB used for training and validation (a) and evaluation (b-f). **a** Three densely annotated ground-truth cubes (5x5x5 µm) located in the calyx brain area were used for training and model validation (Cremi training dataset). In this dataset, synaptic partners are annotated as individual pairs of pre- and post-synaptic sites (purple nodes, pre-synaptic sites marked with a black dot), connected by an edge. The white box in the middle image corresponds to the field of view (848x848x1440 nm) of our U-Net. Every visible feature in this field of view can be used by the network to make the prediction for the single voxel marked in the middle as a white circle. **b** Four sparsely annotated ground-truth datasets located in the lateral horn (peach), calyx (dark blue), ellipsoid body (light blue), protocerebral bridge (green) were used for evaluation. **c**-**f** Example synapse annotations for each of the four datasets from (b). Annotations are sparse and only complete for specific neurons (marked with a star). Pre- and post-synaptic locations are more distant than in (a). **c** Incoming synapses of a Kenyon cell in calyx. **d** Outgoing synapses of a projection neuron in lateral horn. **e** Incoming- and outgoing synapses of a neuron in ellipsoid body. **f** Incoming synapse of a neuron in protocerebral bridge. Rendering in a,b was done with with Catmaid-to-blender [12].



		without neuron segmentation				with neuron segmentation			
		small		big		small		big	
		MSE	CE	MSE	CE	MSE	CE	MSE	CE
standard	MT_2	▲ 0.68	0.69	0.7	0.71	▲ 0.73	0.72	0.74	0.75
	ST	0.69	0.68	0.7	0.72	0.74	0.73	0.74	0.74
curriculum	MT_2	0.69	0.68	0.72	0.73	0.74	0.72	0.76	0.76
	ST	● 0.7	0.7	0.72	◆ 0.74	● 0.76	0.74	0.76	◆ 0.76

Supplementary Figure 2: Validation results on Cremi dataset. **a** Grid-search results in terms of best f-score (higher/darker is better) over all extraction thresholds for various parameter configurations: network size (small: $f = 4$, big: $f = 12$), training loss for $\hat{m}$ (MSE: mean squared error, CE: cross-entropy), balancing strategies (curriculum: $p_{\text{rej}} = 0.95$ until 90k training iterations then $p_{\text{rej}} = 0$, standard: $p_{\text{rej}} = 0.95$ constant), U-Net architectures (ST, MT_2). Left side shows results without post-processing using a neuron segmentation. Glyphs show highest f-score. **b** Precision-recall curves over θ_{CS} for three selected models in (a), respective best f-score for each curve is highlighted.



Supplementary Figure 3: Synaptic partner representation and corresponding CNN architectures. **a** Appearance of a polyadic synapse in EM data with vesicles (black arrow), T-Bar (white arrow), and synaptic cleft (red arrow). **b** Given ground-truth annotations (s, t) for synaptic partner sites, we generate a post-synaptic site mask m (indicating the location of post-synaptic sites) and a field of 3D vectors d (pointing to the corresponding pre-synaptic site). Created spheres around point annotations shown in peach for m and in light blue for d. **c** We train a CNN on m and d to predict $\hat{m}$ and $\hat{d}$, from which we extract synaptic connections. 3D vectors in d and $\hat{d}$ are RGB-color encoded. **d** Investigated CNNs: ST (one U-Net per $\hat{m}$ and $\hat{d}$), MT_1 (one U-Net for both $\hat{m}$ and $\hat{d}$), and MT_2 (one U-Net with two separate upsampling paths per $\hat{m}$ and $\hat{d}$).

Supplementary Notes 1 (Datasets)

We refer to the datasets of manually placed skeleton and synapse annotations (created with Catmaid [13]) that were used to evaluate whole-brain predictions as InCalyx, OutLH, InOutEB, InOutPB, PairCalyx and PairLH. These datasets originate from four different brain regions: the calyx, the lateral horn (LH), the ellipsoid body (EB), and the protocerebral bridge (PB); see Supplementary Figure 1b for a rendering of their location within the full brain. Among those datasets, we distinguish two kinds of annotations: The first four datasets (InCalyx, OutLH, InOutEB, and InOutPB) contain manual skeleton traces of neurons that have all their incoming and/or outgoing synaptic connections annotated, possibly restricted to a well defined region (as reflected in the dataset name, *e.g.*, InCalyx refers to skeletons with all incoming synaptic connections annotated within the calyx). We refer to this kind of dataset as *synapse complete*. The remaining two datasets (PairCalyx, PairLH) consists of skeleton traces and synapse annotations of two sets of neurons, with known number of synaptic connections between every pair of neurons from one set to the other. We refer to this kind of dataset as *pair complete*. Statistics about all six datasets are shown in Supplementary Notes Table 1.

InCalyx contains 528 Kenyon cells from [22] for which inside the calyx all 59,155 input connections have been annotated. A second dataset PairCalyx contains the same 528 Kenyon cells receiving input from 138 olfactory projection neurons from [21] (44,657 connections).

OutLH contains three olfactory projection neurons from [8], for which all 11,429 outgoing connections have been annotated inside the lateral horn. A second dataset PairLH contains 389 neurons from [1] (including the three projection neurons from OutLH) with complete known connectivity inside the lateral horn (24,846 connections).

InOutEB contains 27 neurons from [17], for which all 61,280 incoming and outgoing connections have been annotated inside the EB.

InOutPB contains the same 27 neurons as in InOutEB but neuron parts with all their 14,779 incoming and outgoing connections are located in PB.

We also densely annotated an additional ten cubes with a side-length of 3 μm each (dataset DenseCubes). These cubes are comprised of a total of 270 μm^3 of neural tissue and contain 2800 synaptic connections. The cubes were chosen to be uniformly distributed in the FAFB volume (see Extended Data Figure 1). We chose two cubes each to be *difficult* (*i.e.*, containing imaging artifacts like registration errors, cubes 4 and 5) and *axo-axonic* (*i.e.*, containing disproportionally more axo-axonic synaptic connections, cubes 6 and 7). We will refer to the remaining six cubes as *normal*.

Dataset	Completeness	Connection count	Neuron count (length [mm])	Brain Region	Source
Cremi	dense	1,965	-	calyx	https://cremi.org
InCalyx	input	59,155	528 (213)	calyx	[22]
OutLH	output	11,429	3 (2)	lateral horn	[1]
InOutEB	input & output	61,280	27 (38)	ellipsoid body	[17]
InOutPB	input & output	14,779	27 (16)	protocerebral bridge	[17]
PairCalyx	connectivity	44,657	138 (34), 528 (213)	calyx	[22]
PairLH	connectivity	24,846	389 (243), 389 (243)	lateral horn	[1]

Supplementary Notes Table 1: Ground-truth datasets in FAFB. Cremi dataset was used for training and validation, other datasets were used for evaluation. Completeness column describes the nature of ground-truth annotation: **dense** means that an entire volume is densely annotated; **input** and/or **output** refer to *synapse complete* neurons for which all input and/or output connections are annotated; **connectivity** describes *pair complete* neurons for which all connections between two sets of neurons are annotated.

ID	Type	Connections	Brain Region
1	normal	261	FB
2	normal	254	WED right
3	normal	380	AVLP right
4	difficult	172	AL (VP1d)
5	difficult	197	GNG
6	axo-axonic	447	LH right
7	axo-axonic	279	LH left
8	normal	269	CA left
9	normal	268	MED left
10	normal	273	MED right

Supplementary Notes Table 2: Individual cubes of the DenseCubes dataset. See also Extended Data Figure 1 for a visualization within the FAFB volume.

Supplementary Notes 2 (Evaluation)

We evaluate the accuracy of predicted synaptic partners in two different ways, corresponding to the kind of ground-truth annotations available: For *synapse-complete* neurons, *i.e.*, neurons that (possibly only within a well defined region) have all their incoming and/or outgoing synaptic partners annotated, we evaluate the precision and recall of all synaptic partners that were mapped to those neurons. This evaluation applies to datasets InCalyx, OutLH, InOutEB, and InOutPB. For pairs of neurons with known connectivity, we evaluate the accuracy of correctly predicting the number of synaptic sites between those neurons (datasets PairCalyx and PairLH).

For both evaluation types, we first map each predicted pre- and post-synaptic site $(\hat{s}, \hat{t}) \in \hat{C}$ to a skeleton $n \in N$ (Methods). We will write $n(\hat{s})$ and $n(\hat{t})$ to refer to the mapped skeleton of $\hat{s}$ and $\hat{t}$, respectively. To guide the mapping, we generate a neuron segmentation locally around the synaptic sites to be matched (processed in 2 µm blocks with additional context of 1 µm to avoid border artifacts), using hierarchical agglomeration [5] favoring oversegmentations on affinity predictions obtained from Local Shape Descriptors [14]. We also find it necessary to include manually traced pre- and post-synaptic sites for mapping. This is essential for partnering neurons for which no neuron skeleton exist. In this case, predicted connections are mapped onto single synaptic site nodes. We exclude connections from downstream analysis when either pre- or post-synaptic site connects to a putative glia cell (zero-valued background in the neuron segmentation). See Supplementary Notes Figure 1 for an example segmentation and mapping, zero-background marked with a white arrowhead.

To evaluate precision and recall on synapse-complete neurons, we use the Cremi evaluation procedure (with skeletons IDs instead of neuron segment IDs) with the following two modifications to account for differences in synapse annotation: First, we increase the matching threshold for pre-synaptic sites from 400 nm to 700 nm. This more permissive threshold was empirically found to be necessary to compensate for the larger variance of pre-synaptic site placement in the one-to-many annotations used in the ground-truth (see Supplementary Figure 1c-f for examples). Second, we do not require a predicted post-synaptic annotation to be within a certain threshold distance to a post-synaptic site in the ground-truth, since the ground-truth annotations do not make use of a dedicated post-synaptic site marker. Instead, pre-synaptic nodes are directly connected to a skeleton node of the post-synaptic neuron, which is potentially far away from the predicted post-synaptic site. In summary, we consider a match between $(\hat{s}, \hat{t})$ and a ground-truth annotation (s, t) possible, if $n(\hat{s}) = n(s)$, $n(\hat{t}) = n(t)$, and $|\hat{s} - s| \leq 700\,\text{nm}$. As in the original Cremi evaluation procedure, we perform a Hungarian matching to find at most one-to-one correspondences between possible matches, minimizing their Euclidean distance.

To measure the accuracy of correctly predicting the number of synaptic connections between pairs of neurons, we directly use the result of mapping predicted synaptic sites to skeletons. Let $w(n_1, n_2)$ be the true number of synaptic partners between neurons n_1 and n_2 and

$$\hat{w}(n_1, n_2) = \left|\left\{(\hat{s}, \hat{t}) \in \hat{C} \mid n(\hat{s}) = n_1 \wedge n(\hat{t}) = n_2\right\}\right|$$

the number of predicted synaptic partners. We will refer to w and $\hat{w}$ as the true and predicted *weight* between neurons. Given a weight threshold γ, we report the accuracy for correctly predicting the presence of an edge. An edge is present between a pair of neurons, if and only if the weight is equal or above γ. However, most neuron pairs in the test set do not have a synaptic connection, which would lead to an unreasonably high accuracy stemming from trivially predictable negatives. Therefore, we limit the accuracy analysis to *relevant* neuron pairs $\{(n_1, n_2) \mid w(n_1, n_2) > 0\}$, but nevertheless count each non-relevant pair (n_1, n_2) for which $\hat{w}(n_1, n_2) \geq \gamma$ as an additional false positive.

For the neuron-proximity baseline we randomly select points within a neuron as post-synaptic sites and obtain corresponding pre-synaptic sites by randomly drawing direction vectors from the predicted synaptic partners in FAFB. We then follow the same methodology described above as for the predicted connections with regards to synapse mapping and evaluation. We consider only connections that link different neurons.

Evaluation on Synapse-Complete Neurons

We use the synaptic cleft predictions from [6] to derive a *cleft score* for each putative pair of synaptic partners by taking the maximal synaptic cleft value along a line from the pre- to the post-synaptic site. We then use the product of the connection score and the cleft score to score and threshold synaptic partners (Figure 2a).

Among the synapse-compete datasets, we observe the highest accuracy for dataset InCalyx, which is proximal to the Cremi datasets we used for training and validation. In fact, the f-score of 0.73 for InCalyx is closest to the result on the validation set (f-score 0.76).

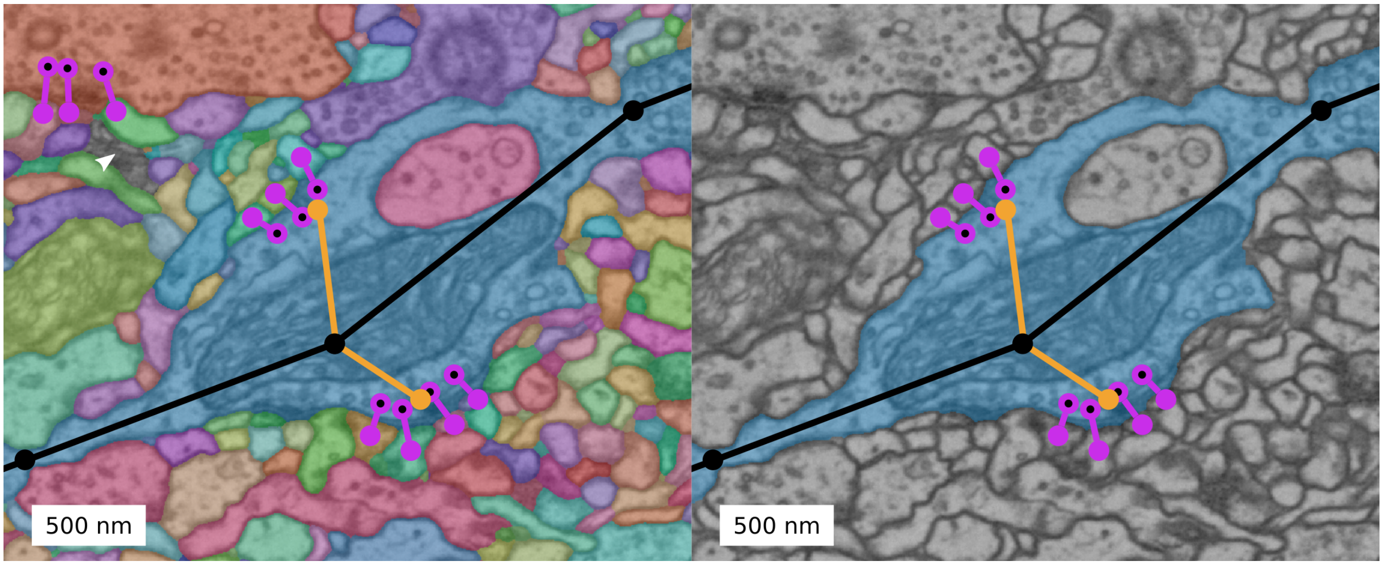
We next studied differences in accuracy based on cell type. Datasets InCalyx and OutLH exclusively contain Kenyon cells and olfactory projection neurons, respectively. The InOutEB and InOutPB datasets contain the same 27 neurons representing columnar neuron types (EP-G, P-EG, P-EN1, P-EN2). Results for performance grouped based on those six cell types are provided in Supplementary Notes Figure 2. For Kenyon cells and olfactory projection neurons, the f-score is evidently the same as for InCalyx and OutLH: **0.73** and **0.68**. For the columnar neuron types both present in the ellipsoid body and the protocerebral bridge, we find best achieved f-scores of **0.66** for EP-G, **0.59** for P-EG, **0.64** for P-EN1, and **0.64** for P-EN2. As there are multiple ways to define a cell type we additionally provide performance per individual neuron (file SupplementaryFile01.csv), which can be used to aggregate statistics for a specific type of neurons[1].

We further investigated the role of a neuron segmentation used for mapping of synapses onto neuron skeletons. For this, we repeated the evaluation of InCalyx with the FFN segmentation from [10] instead of using the segmentation from [14]. Results are shown in Supplementary Notes Figure 3. We observe a decrease in performance from a f-score of 0.73 for Lsd to 0.69 for FFN. We further find that across the entire FAFB dataset, 17% of synaptic partners (either pre-synaptic site or post-synaptic site) are assigned to “background” in the current FAFB segmentation (version: 20200412).

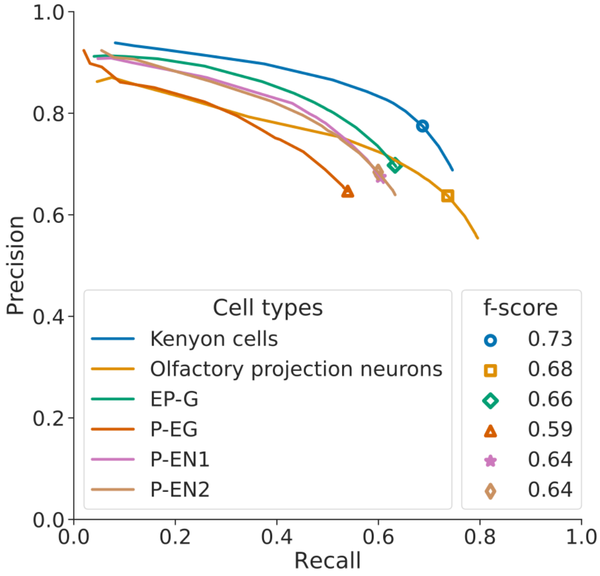
Evaluation on dense cubes

Precision, recall, and f-scores on the DenseCubes dataset are generally consistent with our analysis on the synapse complete datasets (Extended Data Figure 1). Notable outliers are cubes 4

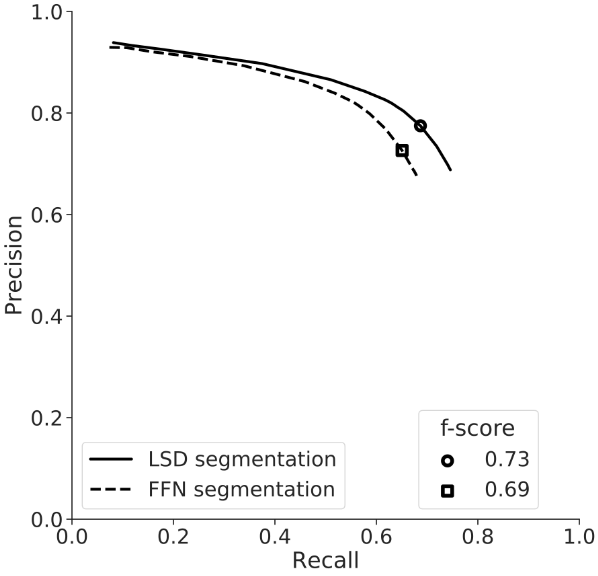
[1] See [17] for a detailed description of cell type nomenclature for the columnar neuron types in the central complex.



Supplementary Notes Figure 1: Mapping of predicted connections to manually created skeletons using neuron segmentation. Left: Around a manually traced skeleton (black line with nodes), we generate a neuron segmentation. Right: Predicted synaptic sites that intersect with the same neuron segment as the skeleton are assigned to that skeleton. For evaluation, we also use manually traced pre-and post-synaptic sites (orange nodes) for mapping. White arrowhead indicates zero-background representing putative thin glial processes.



Supplementary Notes Figure 2: Synaptic prediction performance, cell type comparison. Precision-recall curves using the Cremi metric for the six different cell types Kenyon cells (calyx), olfactory projection neurons (lateral horn), columnar neuron types EP-G, P-EG, P-EN1, P-EN2 (ellipsoid body and protocerebral bridge) over different prediction score thresholds.



Supplementary Notes Figure 3: Impact of different neuron segmentation used for synapse mapping onto neuron skeletons. Precision-recall curves using the Cremi metric for the same set of synapses in calyx (dataset InCalyx) mapped onto skeletons with our original Lsd segmentation versus the FFN segmentation.

and 5 (f-scores of 0.58 and 0.69, respectively) and cubes 9 and 10 (f-scores of 0.69 and 0.68, respectively). Cubes 4 and 5 have been picked to be *difficult*, *i.e.*, they contain imaging artifact like registration errors. Cubes 9 and 10 belong to the medulla (left and right) and the decrease of performance might be attributable to morphological differences that were not encountered during training. The remaining cubes achieve f-scores between 0.70 and 0.79.

We selected three of the *normal* cubes (1, 2, and 3) to be reconstructed independently by two connectome annotators. From those annotations, we computed the inter-human accuracy using the exact same procedure we used for evaluation of our method. For that, we considered one of the two manual reconstructions the ground-truth and computed precision, recall, and f-score of the other one, following the evaluation procedure in [20].

Results are shown in Extended Data Figure 1, top row. The inter-human accuracy varies between 0.73 and 0.83, suggesting that a large part of individual synaptic connections is indeed ambiguous and that it will be unlikely for automatic methods to exceed those scores.

We conducted a comparison of [3] and the current method on the DenseCubes dataset. Results are presented in the same Extended Data Figure 1 and confirm that our current method is consistently more accurate.

Evaluation on neuron connectivity

We make use of the ground-truth datasets PairCalyx and PairLH to evaluate our method in the context of automatically inferring a connectome.

The number of detected synapses depends on the score threshold θ_{CS}. In order to obtain θ_{CS}, we split neurons in PairCalyx and PairLH into a validation and test set. We use θ_{CS} that optimizes the f-score on the validation set and only use neurons in the test set for connectivity analysis. For PairCalyx, we use 105 Kenyon cells as validation set obtaining $\theta_{\text{CS}} = 60$ (10 for synaptic cleft score). Our connectivity test set consists of 138 projection neurons partnering with the remaining 423 Kenyon cells. For PairLH we use all three projection neurons in OutLH and obtain $\theta_{\text{CS}} = 60$ (30 for synaptic cleft score). Our test set consists of the remaining 386 x 386 neurons.

When defining connectivity with $\gamma = 5$, PairCalyx contains 2,039 pairs of neurons that are connected ($w \geq 5$) and 240 that are disconnected $w < 5$, and PairLH contains 1,221 pairs of neurons that are connected and 3,268 that are disconnected. We find that for both datasets, the neuron-proximity baseline has a substantially lower edge accuracy score of **0.39** (−0.57) for PairCalyx and of **0.69** (−0.23) for PairLH (Figure 2e). Note that despite an decrease in precision and recall, accuracy increases due to the contributions of true negatives, which are not counted in precision and recall.

Evaluation on vertebrate neural tissue

To investigate whether the method proposed here generalizes to neural tissue other than the investigated *Drosophila* dataset used here, we trained and evaluated its performance on a mouse cerebellar dataset (dataset MouseCereb). For that, we annotated twelve ground-truth cubes of various regions across the granule cell layer, the Purkinje cell layer, and the molecular layer, imaged at a $4 \times 4 \times 40$nm resolution. Each cube has a side length of 4 µm (∼ 100M voxels). We used seven cubes for training and five for evaluation. Network architecture, training, and evaluation methodology are as described previously with some minor differences: For the network architecture, we used the “small” network size and a cross-entropy loss. We downsampled the volumes by a factor of four in the x- and y-dimension for an effective resolution of $16 \times 16 \times 40$nm prior to training. Consequently, we reduced the downsample factors of the U-Net to $(2, 2, 1)$, $(2, 2, 1)$, and $(2, 2, 3)$. Instead of predicting the location of the post-synaptic partner, we trained the mask network to predict the midpoint at the center of the synapse cleft, and the direction network to predict the direction to the pre-synaptic partner in order to better distinguish the cases of one-to-many and many-to-one synapses which happen frequently in the cerebellar cortex. For evaluation, as cerebellar synapses are much larger than in the fly, we increased the matching distance to 500nm.

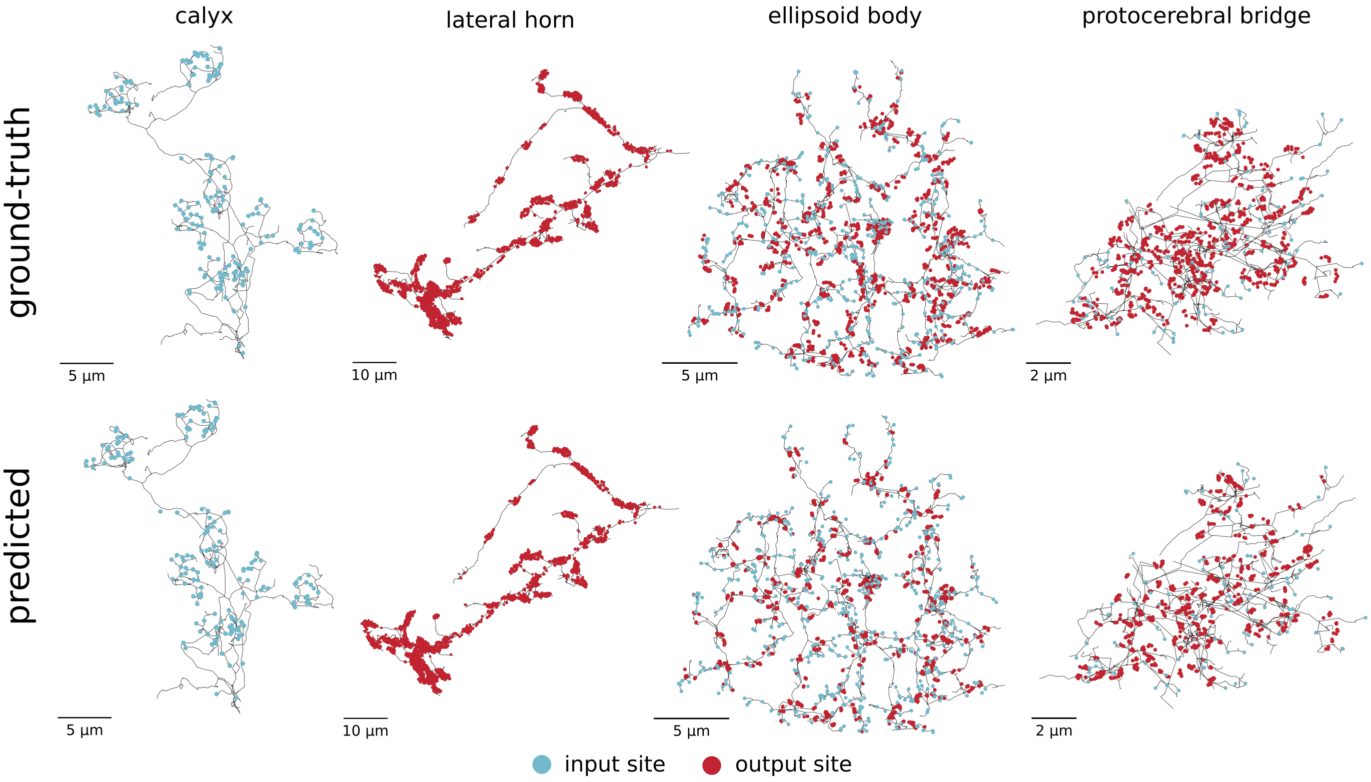
Evaluation results on the five evaluation cubes are shown in Supplementary Notes Figure 6. On this dataset, our method reaches a very high accuracy (f-score 0.94), which might well be attributed to the fact that synapses in vertebrates are larger and less ambiguous.

Error correction using a neuron segmentation

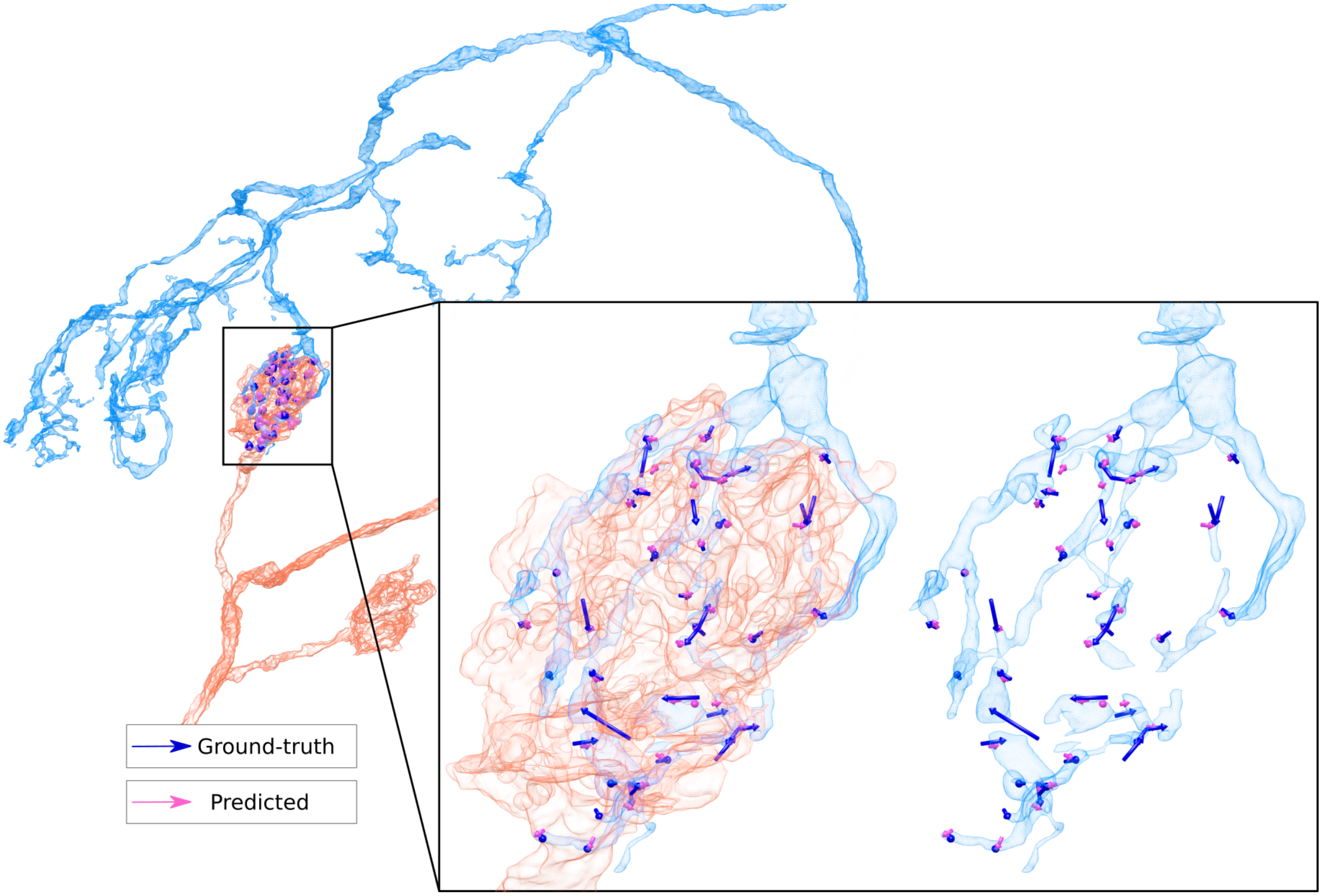
Although a neuron segmentation is not needed for prediction, it can be used to improve accuracy of synaptic partner detection by filtering two types of false positives during post-processing: (1) false positives connecting the same neurite, and (2) duplicate close-by detections of a single synaptic partner pair across the same cleft. The ability to remove duplicate detections around a cleft allows using local non-max suppression (NMS) to identify post-synaptic sites, which naturally gives rise to multiple detections in larger post-synaptic neurites.

We test the impact of using NMS (radius: 40, 80, or 120 nm) for post-synaptic site detection compared to finding connected components (CC) (Methods and Supplementary Figure 2). Overall, scores improve (best f-score 0.76 compared to best f-score without neuron segmentation 0.74). Remarkably, smaller models (∼90% less parameters) improve considerably to the point of matching the best observed score (small ST with curriculum learning using MSE improves f-score from 0.7 to 0.76). For each investigated combination of hyperparameters, NMS performed better than CC.

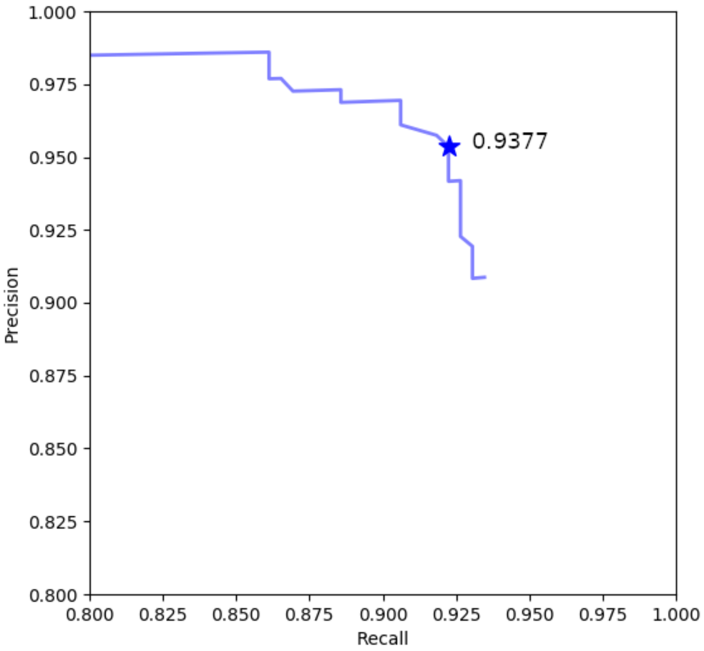
These findings suggest that in the presence of an accurate neuron segmentation smaller, more efficient architectures are on par with larger ones. It should be noted, however, that the neuron segmentation used here for validation was considerably proof-read and that the gain from purely automatic segmentations is likely less.



Supplementary Notes Figure 4: Qualitative results on synapse complete neurons in whole-brain dataset. Example neuron for each of the four brain regions calyx, lateral horn, ellipsoid body and protocerebral bridge with ground-truth (top row) and predicted (bottom row) synaptic connections. We only display neuron parts and connections used for evaluation, *i.e.*, incoming connections in first column and outgoing connections in second column are omitted and neurons are intersected with respective brain regions.



Supplementary Notes Figure 5: Qualitative example of one edge in the connectome. Connection is formed between a projection neuron (red) and a Kenyon cell (blue) with a ground-truth synapse count of 41 (blue arrows) and a predicted of 35 (pink arrows).



Supplementary Notes Figure 6: Evaluation results on dataset MouseCereb. Highlight shows best f-score.

Supplementary Notes 3 (Polarity Analysis)

To test the synapse predictions obtained with our method in a biological context, we used previously published FAFB olfactory projection neurons (PNs) [1], which connect the first order olfactory neuropil, the antennal lobe (AL), with higher order olfactory neuropils, such as the mushroom body or the lateral horn. PNs come in two broad flavours, which depend on whether they receive input in a single (*i.e.*, uniglomerular) or multiple (*i.e.*, multiglomerular) glomeruli of the AL. Uniglomerular PNs in particular have been extensively studied in the past and are thought to be highly polarised, *i.e.*, their dendritic neurites in the AL are predominantly post-synaptic whereas their axons are predominantly pre-synaptic [19, 7, 2, 1]. This is less clear for multiglomerular PNs and there is indication that some of them might be less polarized than their morphologies suggest [1].

To address this, we first sought to computationally split the PNs into axon dendrites using a previously published algorithm based on synaptic flow [1, 13] (Extended Data Figure 2a). In brief, for each of the 346 PNs, we combined semi-manual reconstructions of PNs [1] with the FFN segmentation of the FAFB data set [10] to collect its pre- and post-synapses. This data was then used to draw a path from every post- to every pre-synapse and assign a synapse flow to each neurite by counting the number of pre→post paths traversing that neurite. The part of a PN with the highest synapse flow (the *linker*) was used to split the PNs into axons and dendrites. Manual review of the splits showed that 278 (84%) match the intuitive expectation based on morphology, 42 (12%) contained minor mistakes such as a small misidentified branches and only 26 (7.5%) contained major errors such as flipped axon dendrite. We suspect that at least some cases with perceived split errors might be due genuine biological ambiguity with respect to whether that neuron has a clear dendritic and axonic compartment.

Compartmentalization of the PNs allowed us to calculate a segregation index (SI) that describes how well pre- and post-synapses separate onto axon versus dendrites, respectively [13]. An SI of 1 indicates perfect segregation with the axon being the sole output and dendrites containing all inputs to a neuron. Conversely, an SI of 0 means that axon and dendrites do not differ in their pre-or post-synapse composition, rendering the distinction (Extended Data Figure 2b). Our analyses show that, on average, uni- and multiglomerular PNs do not differ with respect to their SI. We do, however, observe a large variance with some PNs being highly polarized while others do not segregate at all. This demonstrates that PNs are a diverse group of neurons and any broad classification is necessarily an oversimplification. More concretely, our findings suggest that while some highly polarized PNs might act as straight forward relay between the AL and other brain regions, other less polarized PNs likely also engage in local processing—*i.e.*, via presynapse on their dendrites or post-synapses on their axons.

Finally, we address some potential issues and caveats encountered during this analysis. We found that the number of dendritic post-synapses recovered, in particular for uniglomerular PN, was lower than expected. For example, uniglomerular PNs for the DM6 glomerulus had previously been reported to have order 1,200 dendritic postsynapses [16]. For the same type of PNs we recover on average only 282 (± 53) dendritic postsynapses. This is likely due to a combination of two factors: first, the dendrites of PNs we used here were mostly reconstructed “to identification”, not to completion. Second, the quality of image registration varies within the FAFB dataset and the antennal lobe is arguably one of the worst regions. This in turn affects the segmentation we used to map synapses onto PNs. To quantitatively assess how sensitive the mapping process is to incomplete data, we took a PN that had been reconstructed to completion [1], iteratively pruned back its neurites and measured the number of synapses we were able to map onto it (Extended Data Figure 2c). Even mild pruning of about 25% neuronal arbor lead to a severe drop in mapped pre- (40%) and post-synapses (50%). This underlines the importances of good neuronal segmentation and reconstruction in making use of the synapse data presented in this study.

Supplementary Discussion

Prediction Accuracy in FAFB

Our results indicate that the proposed method reliably detects connectome edges with five or more synaptic connections. Qualitatively, we find those results confirmed across different neuron types from all four evaluated datasets: As shown in Supplementary Notes Figure 4, the overall distribution of synaptic sites along the skeleton is generally preserved. In particular, the distribution of predicted pre- and post-synaptic sites of more complex neuron morphologies in the ellipsoid body and the protocerebral bridge agree with the ground-truth distribution.

The high accuracy of detecting connectome edges in dataset PairCalyx and qualitative impressions seem to contrast the comparatively low f-scores obtained on the synapse-complete dataset InCalyx on the same brain region (0.73 without using a cleft prediction and 0.75 with a cleft prediction). For comparison, the example neuron of the calyx shown in Supplementary Notes Figure 4 (first column) has a per-neuron f-score of 0.79, despite the fact that prediction and ground-truth largely agree.

We attribute this discrepancy largely to the fact that the Cremi metric we use to evaluate the synapse-complete datasets is quite conservative (for comparison, also the current leader of the Cremi challenge has a comparatively low f-score of 0.58): due to the Hungarian matching performed between true and predicted synaptic pairs, a predicted connection where at least one of the synaptic sites is slightly incorrectly placed such that it ends up on a different neuron segment adds both to the number of false positives and false negatives. The edge accuracy, on the other hand, would at most count one missing connection (which might even be compensated for by an additional detection further away). As such, the Cremi metric is more sensitive to errors in the neuron segmentation (especially in proximity to synaptic clefts) and spurious or missing manual annotations in ambiguous situations (see Figure 2b, first row left and middle for questionable false positives and false negatives). Furthermore, false merges in the neuron segmentation cause correctly predicted synaptic connections to be incorrectly mapped onto a skeleton, and thus contribute to the number of false positives of the evaluated skeleton. This type of error will be counted in the Cremi metric for each neuron, but in the edge accuracy only if the merger occurs between a pair of neurons contained in the evaluation dataset. Similarly, since the edge accuracy is limited to pairs of neurons with known connectivity, false positives to other neurons are not counted either.

Hence, the Cremi metric should be interpreted as a measure of the whole pipeline accuracy, *i.e.*, not just of the prediction of synaptic partners, but also of the neuron segmentation, of the exact placement of skeleton nodes, and of the mapping of synaptic partners to skeletons. Given a perfect neuron segmentation, the true precision and recall values of the synaptic partner prediction alone would therefore be higher than reported.

Consequently, the overall accuracy of the proposed method will benefit from more accurate neuron segmentations, in particular within the proximity of synaptic sites. Since the prediction of synaptic partners does not require a neuron segmentation (a segmentation is only needed for the mapping of synaptic sites to skeletons), the segmentation can be replaced with a more accurate one in the future to improve the mapping, without having to retrain or reprocess the synaptic partner predictions.

Despite the fact that the absolute value of the Cremi metric is a lower bound to the actual synapse prediction accuracy, we observe a substantial decrease in f-score in brain regions farther away from the calyx. This effect is most notable in terms of increased false negatives in the protocerebral bridge (InOutPB f-score: 0.59), which anecdotally stem from failures to detect axo-axonic links (via manual inspection, see Figure 2b bottom row middle for an example). Less pronounced but still noticeable, we also observe a decrease in performance when comparing the edge accuracy obtained in calyx compared to lateral horn (0.96 versus 0.92). Given that all the training data used was cropped from the calyx, the decrease in performance suggests that the phenotype of synapses is not uniform across neuron types and that the network learnt to recognize neuron type-specific features.

Our observations have two implications for future work: First, training and validation should be carried out on more diverse datasets that capture more of the variation of synapse phenotypes throughout the whole brain. In particular, connectivity accuracy should be manually validated for regions and neuron types that were not part of the training dataset and error estimates should be incorporated into downstream analysis. Second, the quality of a neuron segmentation should be evaluated not just based on overall topological correctness, but also on accuracy close to boundaries, in particular in the proximity of synaptic clefts. Neither of the two most commonly used metrics to evaluate neuron segmentations (expected run length and variation of information) are sensitive to small errors close to synaptic terminals [11].

Model validation on Cremi

We show that the multi-task network MT_2 (two separate upsampling paths for $\hat{m}$ and $\hat{d}$) achieves similar performance compared to two having separate networks (ST), despite having 40 % less parameters. This is somewhat surprising, since the baseline multi-task network MT_1 (single U-Net with two outputs for $\hat{m}$ and $\hat{d}$) failed to converge either for $\hat{m}$ or $\hat{d}$. A similar observation has been made by [18], who independently to our work compared models similar to our ST, MT_1 and MT_2 on a different computer vision task and also found MT_2 (called Y-Net in their work) to perform best. Further improvements in training multi-task networks are possible by exploring different weighting schemes [9, 18] other than simply summing or multiplying both losses as we did here. Already, MT_2 offers a promising starting point for a general volumetric EM U-Net, where multiple different tasks are jointly solved (*e.g.*, the segmentation of boundaries, intracellular organelles, synapses).

Due to the sparsity of synaptic sites, we found that rejection of mini-batches that do not contain synapses is important to prevent the neural networks from converging to trivial solutions (*i.e.*, predicting zero for the post-synaptic site mask). If we train with constant $p_{\text{rej}} = 0$ (no mini-batches are rejected during the entire training), success or failure highly depends on the strength of voxel-wise balancing. We generally observed stable training when we reject empty mini-batches with a probability of $p_{\text{rej}} = 0.95$ (Methods). Furthermore, we found overall better validation performance when using a "curriculum" strategy, *i.e.*, we first train with a rejection probability of $p_{\text{rej}} = 0.95$ until 90k iterations and $p_{\text{rej}} = 0$ afterwards. The average f-score of all 16 tested setups for a constant rejection probability is 0.71 ± 0.02 compared to 0.73 ± 0.02 for the curriculum strategy (see f-score results in top two rows and bottom two rows in Supplementary

Figure 2). A likely explanation for the increase in accuracy is that with a constant rejection probability of $p_{\text{rej}} = 0.95$, mini-batches containing synapses are overrepresented during training and the resulting networks are weaker in correctly classifying negatives in areas that do not contain synapses such as the cell interior of large neurons.

Role of Neuron Segmentation

Not requiring a neuron segmentation for prediction has two immediate advantages: First, training data can be generated much quicker, since only synaptic partners have to be annotated (two points connected by a line for each synaptic connection). Second, neuron segmentations are subject to change (at the time of writing, there are three different versions of neuron segmentations for FAFB).

Since an updated neuron segmentation changes only the mapping of already found synaptic partners onto neurons, synaptic partners do not have to be detected again. Furthermore, the results of our method can easily be incorporated into a segmentation proof-reading workflow: As errors in the segmentation are fixed (through splits and merges), the mapping of synaptic partners can be updated on the fly. Independently predicted synaptic partners might even be useful to localize errors in a neuron segmentation.

Our method can potentially find synaptic partners that a neuron-segmentation-based method could not find: if two neurons were wrongly merged in a segmentation, synapses between the two neurons would be "invisible" to a method that relies on an accurate segmentation.

Nevertheless, a neuron segmentation is still needed to map predicted synaptic partners onto neurons. This is a potential source of errors, also in our pipeline. Some neuron segmentation methods have inaccuracies in the vicinity of synapses: 17% of synaptic sites are assigned to "background" under the current FFN segmentation for FAFB (version: 20200412).

For the evaluation presented here, a segmentation was needed to map synaptic partners to neurons. Consequently, some of the errors reported here might be due to errors in the segmentation.

Limitations

Our model is well-suited to detect one-to-many synapses (one pre-synaptic site targeting many post-synaptic sites), as it explicitly extracts an individual connection for each detected post-synaptic site. It is however less suited to detect the relatively rare many-to-one synapses, which can be found for instance in the α-lobe of the mushroom body [15]. If we use connected components to find post-synaptic sites in $\hat{m}$, we are likely to only find a single site for a many-to-one synapse and thus would also only extract a single connection.

Finally, it should be taken into account that there is uncertainty about what kind of functional features can be inferred from ultra-structure alone. Although recent work shows promise in relating, *e.g.*, neurotransmitter identities to morphological features visible in EM [4], connectomics data will have to be complemented by physiological experiments and molecular information to gain insights into functional features.

References

[1] Alexander Shakeel Bates, Philipp Schlegel, Ruairí JV Roberts, Nikolas Drummond, Imaan FM Tamimi, Robert Gillies Turnbull, Xincheng Zhao, Elizabeth C Marin, Patricia Demetria Popovici, Serene Dhawan, et al. Complete connectomic reconstruction of olfactory projection neurons in the fly brain. *bioRxiv*, 2020.

[2] Matthew E Berck, Avinash Khandelwal, Lindsey Claus, Luis Hernandez-Nunez, Guangwei Si, Christopher J Tabone, Feng Li, James W Truman, Rick D Fetter, Matthieu Louis, Aravinthan DT Samuel, and Albert Cardona. The wiring diagram of a glomerular olfactory system. *eLife*, 5:e14859, May 2016. ISSN 2050-084X. doi: 10.7554/eLife.14859.

[3] Julia Buhmann, Renate Krause, Rodrigo Ceballos Lentini, Nils Eckstein, Matthew Cook, Srinivas Turaga, and Jan Funke. Synaptic partner prediction from point annotations in insect brains. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 309–316. Springer, 2018.

[4] Nils Eckstein, Alexander S. Bates, Michelle Du, Volker Hartenstein, Gregory S.X.E. Jefferis, and Jan Funke. Neurotransmitter classification from electron microscopy images at synaptic sites in drosophila. *bioRxiv*, 2020. doi: 10.1101/2020.06.12.148775. URL https://www.biorxiv.org/content/early/2020/06/13/2020.06.12.148775.

[5] Jan Funke, Fabian David Tschopp, William Grisaitis, Arlo Sheridan, Chandan Singh, Stephan Saalfeld, and Srinivas Turaga. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1669–1680, 05 2018.

[6] Larissa Heinrich, Jan Funke, Constantin Pape, Juan Nunez-Iglesias, and Stephan Saalfeld. Synaptic cleft segmentation in non-isotropic volume electron microscopy of the complete drosophila brain. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 317–325. Springer, 2018.

[7] Jane Anne Horne, Carlie Langille, Sari McLin, Meagan Wiederman, Zhiyuan Lu, C Shan Xu, Stephen M Plaza, Louis K Scheffer, Harald F Hess, and Ian A Meinertzhagen. A resource for the *Drosophila* antennal lobe provided by the connectome of glomerulus VA1v. *eLife*, 7:e37550, November 2018. ISSN 2050-084X. doi: 10.7554/eLife.37550.

[8] Paavo Huoviala, Michael-John Dolan, Fiona Love, Shahar Frechter, Ruairi JV Roberts, Zane Mitrevica, Philipp Schlegel, Alexander Shakeel Shakeel Bates, Yoshinori Aso, Tiago Rodrigues, et al. Neural circuit basis of aversive odour processing in drosophila from sensory input to descending output. *bioRxiv*, 2018.

[9] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

[10] Peter H. Li, Larry F. Lindsey, Michał Januszewski, Zhihao Zheng, Alexander Shakeel Bates, István Taisz, Mike Tyka, Matthew Nichols, Feng Li, Eric Perlman, Jeremy Maitin-Shepard, Tim Blakely, Laramie Leavitt, Gregory S. X. E. Jefferis, Davi Bock, and Viren Jain. Automated Reconstruction of a Serial-Section EM Drosophila Brain with Flood-Filling Networks and Local Realignment. *bioRxiv*, April 2019. doi: 10.1101/605634.

[11] Stephen M Plaza and Jan Funke. Analyzing image segmentation for connectomics. *Frontiers in Neural Circuits*, 12, 2018.

[12] Philipp Schlegel, Michael J Texada, Anton Miroschnikow, Andreas Schoofs, Sebastian Hückesfeld, Marc Peters, Casey M Schneider-Mizell, Haluk Lacin, Feng Li, Richard D Fetter, et al. Synaptic transmission parallels neuromodulation in a central food-intake circuit. *Elife*, 5:e16799, 2016.

[13] Casey M Schneider-Mizell, Stephan Gerhard, Mark Longair, Tom Kazimiers, Feng Li, Maarten F Zwart, Andrew Champion, Frank M Midgley, Richard D Fetter, Stephan Saalfeld, et al. Quantitative neuroanatomy for connectomics in drosophila. *Elife*, 5:e12059, 2016.

[14] Arlo Sheridan, Tri Nguyen, Diptodip Deb, Wei-Chung Allen Lee, Stephan Saalfeld, Srini Turaga, Uri Manor, and Jan Funke. Local shape descriptors for neuron segmentation. *bioRxiv*, 2021. doi: 10.1101/2021.01.18.427039. URL https://www.biorxiv.org/content/early/2021/01/18/2021.01.18.427039.

[15] Shin-ya Takemura, Yoshinori Aso, Toshihide Hige, Allan Wong, Zhiyuan Lu, C Shan Xu, Patricia K Rivlin, Harald Hess, Ting Zhao, Toufiq Parag, et al. A connectome of a learning and memory center in the adult drosophila brain. *Elife*, 6:e26975, 2017.

[16] William F Tobin, Rachel I Wilson, and Wei-Chung Allen Lee. Wiring variations that enable and constrain neural computation in a sensory microcircuit. *eLife*, 6:e24838, May 2017. ISSN 2050-084X. doi: 10.7554/eLife.24838.

[17] Daniel B Turner-Evans, Kristopher T Jensen, Saba Ali, Tyler Paterson, Arlo Sheridan, Robert P Ray, Tanya Wolff, J Scott Lauritzen, Gerald M Rubin, Davi D Bock, et al. The neuroanatomical ultrastructure and function of a biological ring attractor. *Neuron*, 108(1):145–163, 2020.

[18] Kaiqiang Wang, Jiazhen Dou, Qian Kemao, Jianglei Di, and Jianlin Zhao. Y-net: a one-to-two deep learning framework for digital holographic reconstruction. *Optics Letters*, 44(19):4765–4768, 2019.

[19] Rachel I. Wilson. Early olfactory processing in drosophila: mechanisms and principles. *Annual review of neuroscience*, pages 217–241, 2013.

[20] C Shan Xu, Michal Januszewski, Zhiyuan Lu, Shin-ya Takemura, Kenneth Hayworth, Gary Huang, Kazunori Shinomiya, Jeremy Maitin-Shepard, David Ackerman, Stuart Berg, et al. A connectome of the adult drosophila central brain. *BioRxiv*, 2020.

[21] Zhihao Zheng, J Scott Lauritzen, Eric Perlman, Camenzind G Robinson, Matthew Nichols, Daniel Milkie, Omar Torrens, John Price, Corey B Fisher, Nadiya Sharifi, et al. A complete electron microscopy volume of the brain of adult drosophila melanogaster. *Cell*, 174(3):730–743, 2018.

[22] Zhihao Zheng, Feng Li, Corey Fisher, Iqbal J Ali, Nadiya Sharifi, Steven Calle-Schuler, Joseph Hsu, Najla Masoodpanah, Lucia Kmecova, Tom Kazimiers, et al. Structured sampling of olfactory input by the fly mushroom body. *bioRxiv*, 2020.