nature portfolio

https://doi.org/10.1038/s41593-021-00895-5

Supplementary information

A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging

In the format provided by the authors and unedited

Supplementary Notes

- Supplementary Note 1 | Noise-matching of resampled ground truth data.
- Supplementary Note 2 | Dependence of performance on hyper-parameters, overfitting and network architecture.
- Supplementary Note 3 | Discrete spikes and single-spike precision.

Supplementary Figures

- Figure S1 | Example of history-dependence of spike-evoked fluorescence.
- Figure S2 | Natural resampling of the ground truth vs. artificially sampled noise.
- Figure S3 | Illustration of error, bias and correlation metrics.
- Figure S4 | Parameter-robustness of CASCADE with respect to hyper-parameters.
- Figure S5 | Comparison across deep learning architectures.
- Figure S6 | Neuropil decontamination improves spike inference.
- Figure S7 | Analysis of single-spike resolution of spike inference.
- Figure S8 | Parameter changes leave mutual predictability across datasets largely unchanged.
- Figure S9 | Predicting cross-dataset predictability with a generalized linear model (GLM).
- Figure S10 | Bias of predictions across spike rates.
- Figure S11 | Spike inference with CASCADE improves pairwise correlations for population imaging.
- Figure S12 | Non-Gaussian ground truth smoothing kernel.
- Figure S13 | Locations of neurons in the dorsal telencephalon of adult zebrafish from ground dataset DS#08.

Supplementary Table

- Supplementary Table 1 | Optimal parameters for the model-based spike detection algorithms.

SUPPLEMENTARY NOTE 1

NOISE-MATCHING OF RESAMPLED GROUND TRUTH DATA

To ensure reliable inference of spike rates it is advantageous to train the supervised deep network with a training dataset that matches the noise level of the test neuron from the population imaging data. In practice, the noise level of each neuron from the population imaging data is determined and an existing model trained with approximately the same noise levels is loaded for spike inference. Noise-suppression is only effective if the noise statistics of the ground truth used for training the network resemble the noise statistics of the calcium data used for testing. To generate a ground truth that matches this requirement, we tested two different approaches to increase the noise of ground truth recordings to match population recordings (Fig. S2).

First, we used the raw imaging data to extract not only the mean fluorescence trace, but the fluorescence trace of each pixel of the region of interest (ROI) that defines the neuron. To achieve a given noise level ν, a random subset of pixels was drawn from the ROI pixels until the average fluorescence trace of this sub-ROI reached the desired noise level (Fig. S2a). This method generates realistic noise characteristics through spatial subsampling but is computationally costly and cannot be applied to ground truth datasets when raw fluorescence movies are not available.

Alternatively, we extracted only the mean fluorescence trace of the ROI of a ground truth neuron and added artificial noise until the overall high-frequency noise matched the test calcium dataset (Fig. S2a). This procedure can in theory be repeated to produce an infinite number of examples (replicas) from a simple ground truth recording. However, since the mean ΔF/F of a ground truth recording already is associated with a certain noise level, these noise patterns would be correlated across replicas. To avoid this undesired effect, which could lead to overfitting of correlated noise during training, the number of replicas was restricted to a number n that was computed with the noise level of the mean ΔF/F of a neuronal ROI, ν_{ROI}, and the target noise level, ν_{target}, by $n = \left(\nu_{\mathrm{target}}/\nu_{\mathrm{ROI}}\right)^2$ and thresholded at a maximum of $\mathrm{n} = 500$. We tested both simple Gaussian noise as well as Poisson noise, where the variance of the noise is proportional to the signal amplitude, as is typical for photon shot noise.

To test whether artificial noise enables the network to learn and suppress natural noise patterns we generated ground truth with either natural (spatial sub-sampling) or artificial (Gaussian or Poisson) noise for a subset of the available ground truth datasets (Fig. S2a-c). Using models trained on artificial rather than natural noise slightly but significantly decreased the correlation of predictions with the ground truth when applied to test datasets based on natural noise (decrease for Gaussian noise: Δ_1 = 0.010 ± 0.004, pseudo-median ± 95% C.I., p < 1e-4, paired Wilcoxon test; decrease for Poisson noise: Δ_1 = 0.006 ± 0.004, p < 0.005). This decrement indicates how much better a model trained with naturally sampled noise would likely perform. Conversely, testing models trained with artificially generated ground truth on artificially instead of naturally sampled ground truth increased the correlation of predictions slightly (Gaussian noise: Δ_2 = 0.006 ± 0.008, p = 0.13; Poisson noise: Δ_2 = 0.005 ± 0.009, p = 0.28). This increment indicates how much the correlations with the ground truth will be overestimated when training and testing is done with artificially generated ground truth only. Differences were generally more pronounced for larger noise levels, but overall remained very small compared to the absolute values (not shown). We therefore conclude that artificial noise allows deep networks to effectively learn noise patterns that can be applied to natural noise, with only minor performance loss compared to spatially subsampled recordings, and we further conclude that using Poisson-distributed noise yields slightly improved performance compared to artificial Gaussian noise.

SUPPLEMENTARY NOTE 2

DEPENDENCE OF PERFORMANCE ON HYPER-PARAMETERS, OVERFITTING AND NETWORK ARCHITECTURE

We tested how spike inference performance depends on the choice of hyper-parameters and network architecture. Networks were trained on a specific ground truth dataset using all neurons except one, which was held out for testing (leave-one-out strategy). The algorithm turned out highly robust with respect to changes of the optimizer for gradient descent, the batch size during learning, the number of convolutional features per layer, the number of neurons in the dense layer, and the extent of the temporal window of the receptive field (Fig. S4a-e). All of these observations were confirmed over a surprisingly large range, indicating that the network performance is highly robust with respect to any hyper-parameter choices.

In addition, we investigated potential overfitting of the training dataset. The performance of the network was high already after one training epoch (*i.e.*, as soon as every sample had been seen once by the network), then reached a maximum after 10-30 training epochs, and slightly decreased thereafter (Fig. S4f). This learning behavior suggests only moderate overfitting. At the same time the training loss decreased monotonically (Fig. S4g). We believe that the high abundance of noise and sparseness of events acts as a natural regularizer that prevents overfitting. More importantly, while the learning curve was smooth on average (Fig. S4f), individual network instances sometimes reached unfavorable states. As expected from known properties of deep networks[77], this effect could be easily eliminated by ensemble averaging over 5 networks (Fig. S4h).

Finally, we also tested the performance when employing a network architecture different from the standard convolutional architecture ('default'). We tested a large variety of standard deep learning architectures, including recurrent LSTM networks and non-convolutional deep networks with only the input, output and loss function remaining unchanged (see Methods for detailed descriptions and explanations). Most of these networks performed well, and the performance of several networks was statistically indistinguishable: the default convolutional network, a convolutional network with reduced filter size, a locally connected network, and a bi-directional LSTM network (Fig. S5a). Much smaller networks (single convolutional layer, Fig. S8a) tended to underfit the ground truth. On the other hand, networks with larger numbers of parameters (the deeper convolutional networks and the locally connected network) overfitted the data when training continued (dashed lines in Fig. S5b), consistent with previous observations[16].

The locally connected network and the bi-directional LSTM network performed equally well compared to the default convolutional network despite very different architectures. However, some architectures that were not adapted to spike inference showed lower performance, for example the naïve LSTM network, which by its recurrent design prevents the network from looking precisely at the time point of interest (see Methods for details). Another example is a network identical to the default convolutional network but with purely linear activation functions (Fig. S5a,b), which prevents the algorithm from non-linearly adjusting decision boundaries.

77. Fort, S., Hu, H. & Lakshminarayanan, B. Deep Ensembles: A Loss Landscape Perspective. Preprint at https://arxiv.org/abs/1912.02757 (2019).

SUPPLEMENTARY NOTE 3

DISCRETE SPIKES AND SINGLE-SPIKE PRECISION

Many existing spike inference methods do not aim at the inference of spike rates, but rather of discrete spikes[11,12,15]. Previous publications reported that the average ΔF/F value triggered by a single spike is larger than zero, and that the calcium transients corresponding to single spikes can be detected in selected neurons[34]. However, the precise identification of individual spikes in practical situations is more challenging since a detection scheme should also work in unseen data. In addition, the task may depend in unknown ways on the variable expression of calcium indicators, on shot noise, on other noise sources, on low sampling rates and on the non-linear response of calcium indicators[3,22,30,31,78]. It is therefore not clear whether discrete spikes can be reliably inferred in realistic scenarios. We therefore devised two approaches to test whether single-spike precision can be achieved. First, we focused on single, isolated spikes in the ground truth and compared them with inferred spike rates for the same time window. This approach is discussed in the main text. Second, we transformed spike rates into discrete spikes and analyzed whether the discretization improved spike inference.

With respect to the second approach, we argue that the spike rates inferred by the deep network will exhibit a tendency to quantize if the predictions are close to single-spike precision. Therefore, a procedure that takes into account the prior about discretized spiking could improve the inferred spike rates. We therefore applied an algorithmic procedure that uses prior knowledge about the spike rate (spiking probability) waveform associated with a single spike to fill up the inferred probability trace with discrete spikes using an optimization procedure based on Monte-Carlo importance sampling (Methods).

We found that spiking probabilities that were almost correctly inferred by the deep network were optimized by suppression of noise or by rounding of close matches (blue arrows in Fig. S7a-d). However, this procedure can also enhance small false positive or small negative errors (red arrows in Fig. S7a-d). Over all ground truth datasets, the correlation metric slightly but consistently decreased when spike rates were discretized (Fig. S7e), indicating that the data quality did not allow for efficient use of the prior. Although discretized spike rates tended to decrease the error (Fig. S7e), this effect was primarily due to suppression of small noise events in the absence of spiking, and we found that this positive effect could be achieved without reduction of the correlation by thresholding the inferred spike rates (Fig. S7d,f). Together, this suggests that the available datasets do not permit discretization of predicted spike rates without performance loss.

Despite these caveats, discretization of spike rates might still be useful for two reasons. First, discrete spike events may be more intuitive visualizations of activity than smooth probabilities. Second, while spike rates are smoothed with a Gaussian kernel for each spike, the detection of a single spike that optimally explains this spike provides better temporal resolution. Here, we see a potentially useful application of discrete spike inference, which is, however, beyond the scope of this study. We include the algorithm to discretize spike rates as a script in the public repository (https://git.io/JtZe4).

78. Beaulieu-Laroche, L., Toloza, E. H. S., Brown, N. J. & Harnett, M. T. Widespread and Highly Correlated Somato-dendritic Activity in Cortical Layer 5 Neurons. *Neuron* **103**, 235-241.e4 (2019).

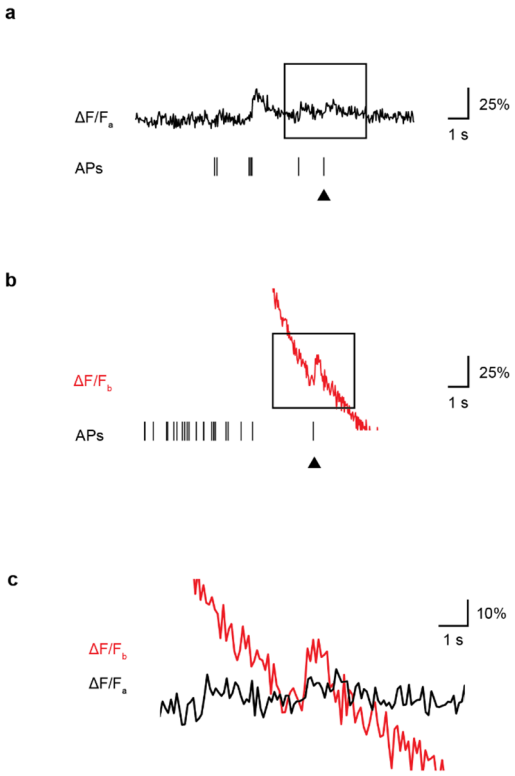

Figure S1 | Example of history-dependence of spike-evoked fluorescence. Calcium recordings of the same neuron during a phase of sparse spiking **(a)** and after a burst **(b)** are shown. Single spikes are marked with a black arrowhead. The corresponding ΔF/F changes for cases (a) and (b) are overlayed and magnified in **(c)**, clearly showing a much larger fluorescence increase evoked by a single spike after the burst. This amplification is due to the non-linear cooperative calcium binding of GCaMP6f, with a larger fraction of indicator molecules being in a pre-bound state briefly after the burst. Example taken from DS#06.

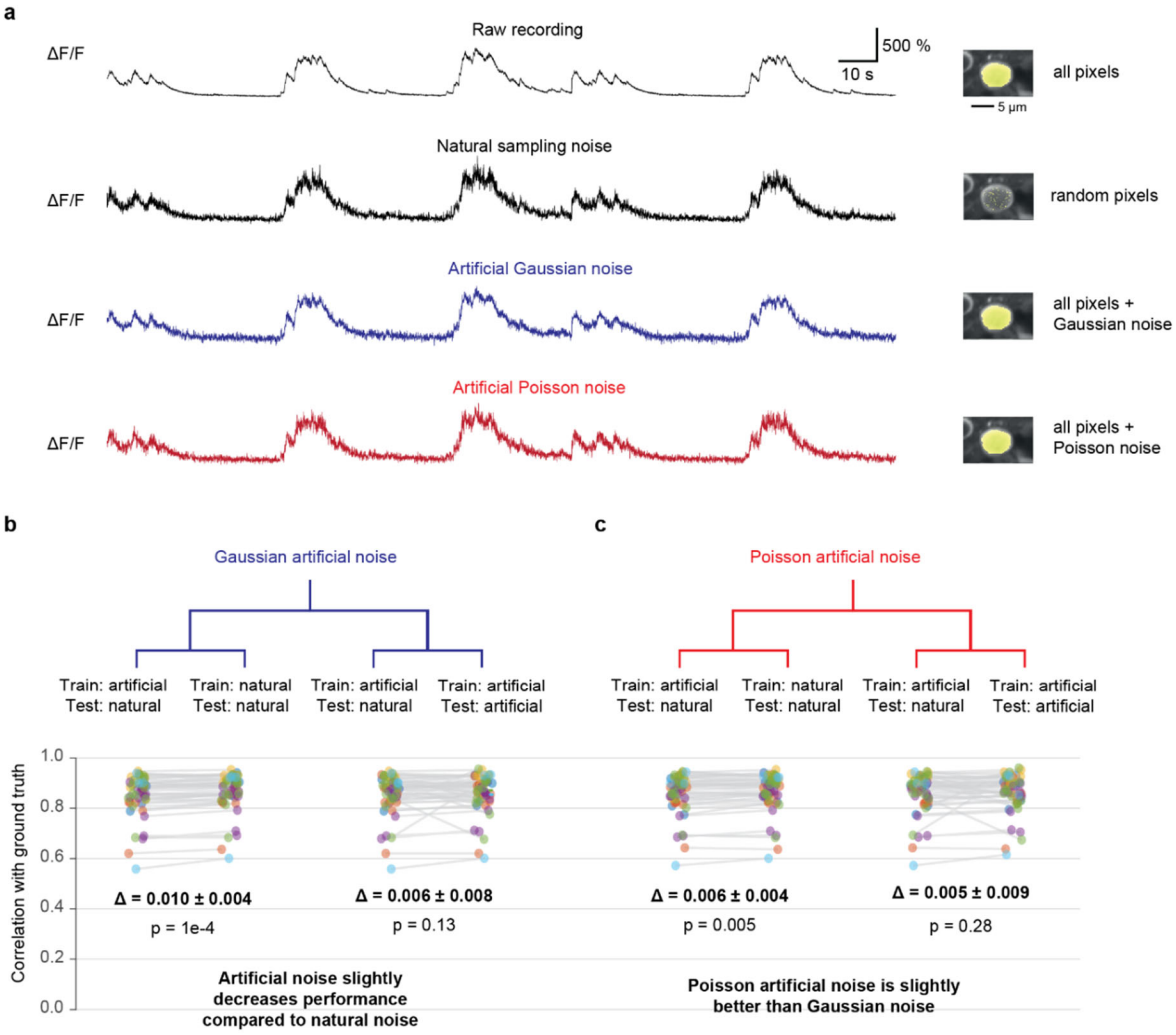

Figure S2 | Natural resampling of the ground truth vs. artificially sampled noise. a, Illustration of the different procedures to generate noisy ground truth recordings from raw ground truth. Row 1: Raw ground truth calcium recording, mean fluorescence. Row 2: Same ground truth recording, spatial subsampling (natural sampling) of random pixels in the ROI. Row 3: Same recording, mean fluorescence with added Gaussian noise. Row 4: Same recording, mean fluorescence with added Poisson noise. **b,** Testing different noise modes. For each condition, two comparisons between models are made: 1) Trained with artificial and tested with natural noise vs. trained and tested with natural noise. This comparison shows how much worse the predictions become when applying the algorithm to naturally sampled calcium recordings while training the algorithm with artificially sampled noise. 2) Trained with artificial and tested with natural noise vs. trained and tested with artificial noise (example with red arrow). This comparison shows how much the procedure of training and testing with artificial noise overestimates the performance compared to the realistic case of training with artificial noise and applying the model to naturally sampled calcium recordings. Each data point represents a neuron, colors indicate ground truth datasets. Differences Δ were computed as pseudo-median ± 95% confidence intervals. P-values and pseudo-medians were computed using two-sided paired Wilcoxon signed-rank tests and are shown in the figure itself. All analyses were performed with a standardized noise level of $\nu = 2$.

Figure S3 | Illustration of error, bias and correlation metrics. As metrics that do not only measure the similarity (correlation) of ground truth spike rates and inferred spike rates, the error and bias also indicate absolute deviations from true spike rates. **a**, Example ΔF/F trace (dark blue), true spike rate (orange) and inferred spike rate prediction (black). The area under the spike rate traces corresponds to the number of true or inferred spikes. The difference between true and predicted spike rates gives false positives (light blue area) and false negatives (red area). **b**, The unsigned integral of the false positives and negatives, divided by the integral of true positives yields the error, while the difference yields the bias. The normalization by the integral of true positives, i.e., the true number of spikes makes errors and biases comparable across spike rate conditions. A side-effect of this normalization is that for neurons that spike only very sparsely, (relative) errors are systematically higher. Correlation is defined as the Pearson correlation coefficient computed with ground truth spike rates and inferred spike rates.

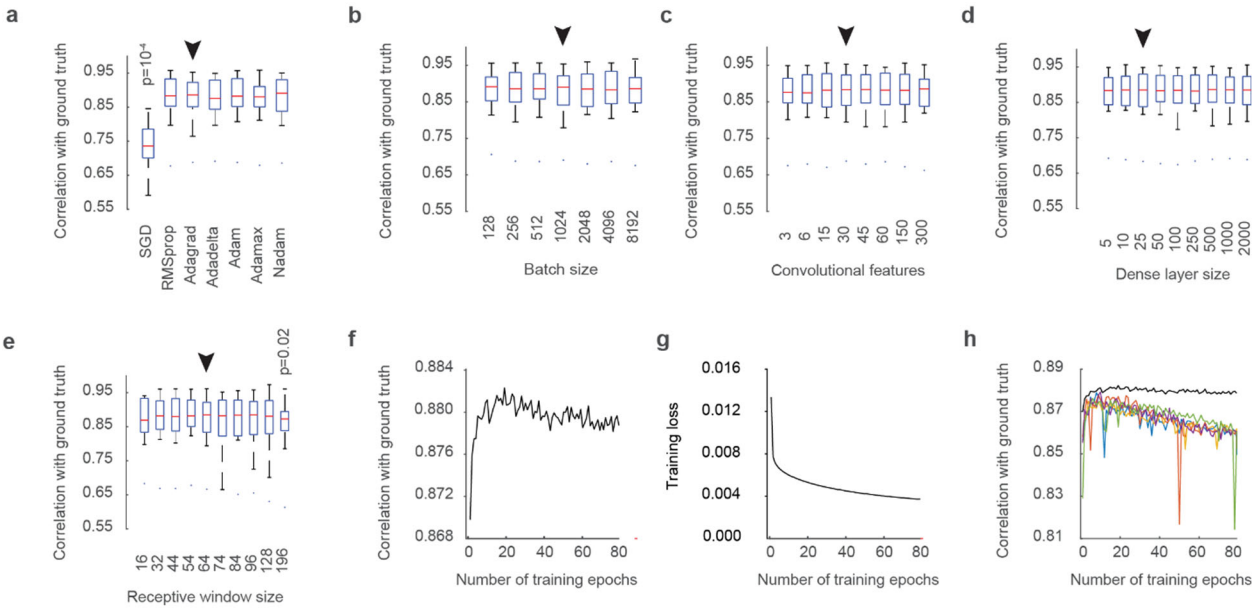

Figure S4 | Parameter-robustness of CASCADE with respect to hyper-parameters. All parameter studies were performed with DS#04 at a calcium imaging rate of 7.5 Hz and a resampled noise level of 2. Box plots reflect a distribution over the n=15 neurons in the dataset. The median is indicated by the central line, 25th and 75th percentiles by the box, and maximum/minimum values excluding outliers (points) by the whiskers. Networks were trained for 10 epochs with all data except for a single neuron, which was used for testing. Ensembles of five networks were used and results were averaged across 3 iterations. Results were compared to the default configuration (arrowhead) using a two-sided Wilcoxon paired signed-rank test. No configuration was significantly different from the default configuration ($p>0.05$), with two exceptions as indicated by p-values above the corresponding boxes. **a,** Performance is robust with respect to choice of the gradient descent optimizer, unless the naive stochastic gradient descent (SGD) is selected ($p = 10^{-4}$). **b,** Performance is robust with respect to the batch size used during learning. Batch sizes can influence the efficiency of gradient descent. **c,** Performance is robust against large variations in the number of features of the convolutional layers. Numbers indicate the mean number of features across the three convolutional layers. **d,** Performance is robust against large variations in the number of neurons in the dense layer. **e,** Performance is robust against large variations in the size of the receptive window. No significant effects were found, except for a decrease with the largest window size ($p = 0.02$). A receptive window of 64 data points corresponds *e.g.* to 64/7.5 ≈ 8.5 s in this setting. **f,** The mean correlation with the ground truth across neurons initially increases and subsequently decreases slightly during training. **g,** The training loss decreases monotonically during training. **h,** While the performance of the network is stable across epochs when using an ensemble of 5 instantiations (black), erroneous and unpredictable deviations can be observed for algorithms based on a single network (colored learning curves).

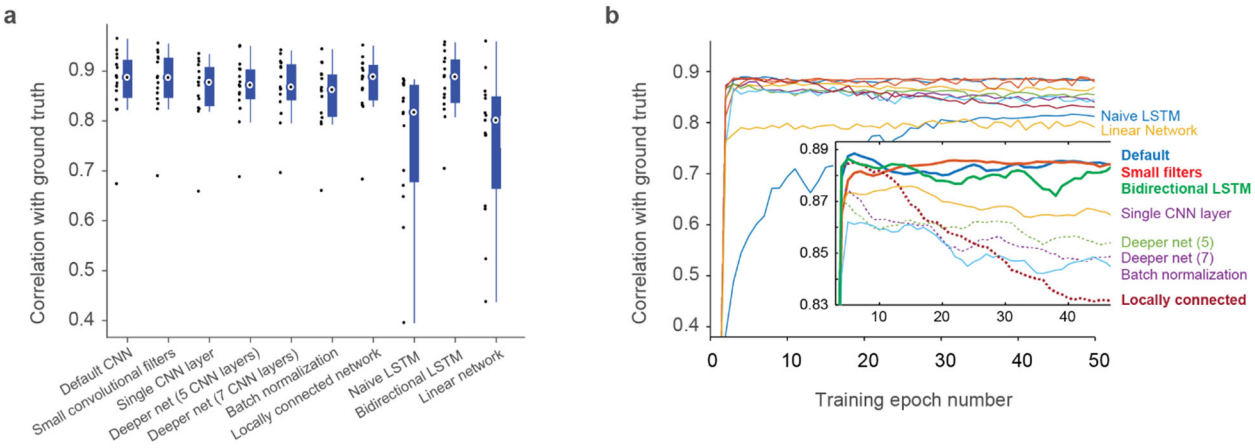

Figure S5 | Comparison across deep learning architectures. a, Comparison of the default network with a network with smaller filters (Δ = -(0.002 ± 0.006), p = 0.28; paired two-sided Wilcoxon signed-rank test; pseudo-median Δ ± 95% confidence interval), with a single layer network (Δ = 0.013 ± 0.009, p = 0.008), with a deeper net (5 layers, Δ = 0.008 ± 0.011, p = 0.09), with a deeper net (7 layers, Δ = 0.008 ± 0.009, p = 0.07), with a network using batch norm (Δ = 0.023 ± 0.010, p = 0.0003), with a locally connected network (Δ = 0.002 ± 0.005, p = 0.33), with a naïve LSTM network (Δ = 0.070 ± 0.063, p = 0.00006), with a bidirectional LSTM network (Δ = 0.002 ± 0.008, p = 0.56) and with a simple linear network (Δ = 0.085 ± 0.080, p = 0.00006). Compared for a single dataset (DS#04), sampled at 7.5 Hz at a standardized noise level of 2. For detailed descriptions of all architectures, see Methods. For box plots, the median is indicated by the central line, 25th and 75th percentiles by the box, and maximum/minimum values excluding outliers (points) by the whiskers. **b,** Learning curves across epochs for all networks. The inset highlights the significant overfitting resulting from relatively large networks (deeper net 5, deeper net 7 and locally connected network; dashed lines).

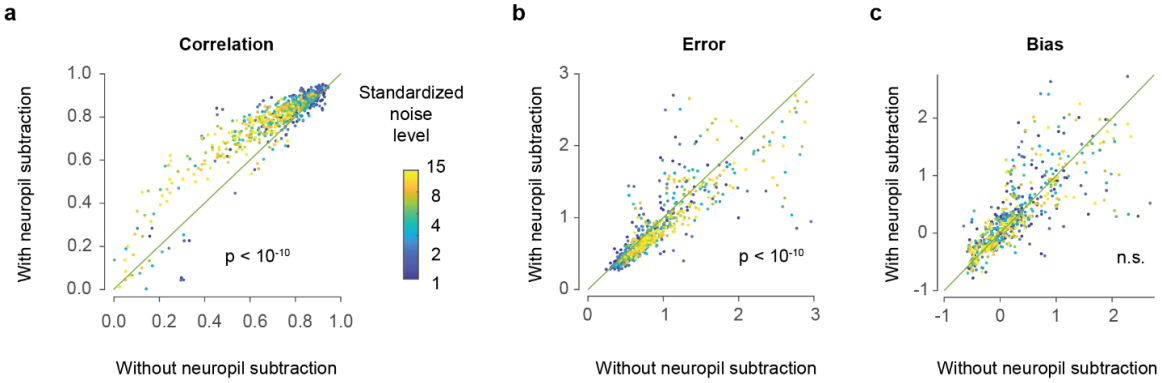

Figure S6 | Neuropil decontamination improves spike inference. For datasets DS#10-13 (Huang et al. (2019), datasets from the Allen Institute), ground truth data were extracted both with and without simple subtractive neuropil decontamination. **a,** The performance (correlation) is improved by neuropil decontamination. The same calcium recordings were analyzed for several noise levels (color-coded). The p-value (paired two-sided signed-rank test) was $<10^{-10}$ for every noise level but decreased for higher noise levels. **b**, Same for the error metric. Errors were significantly reduced after neuropil decontamination. **c**, Same for the bias metric. No significant effect of neuropil decontamination was observed. Color-coding indicates the standardized noise level ν (in $\% \cdot Hz^{-1/2}$) for each data point.

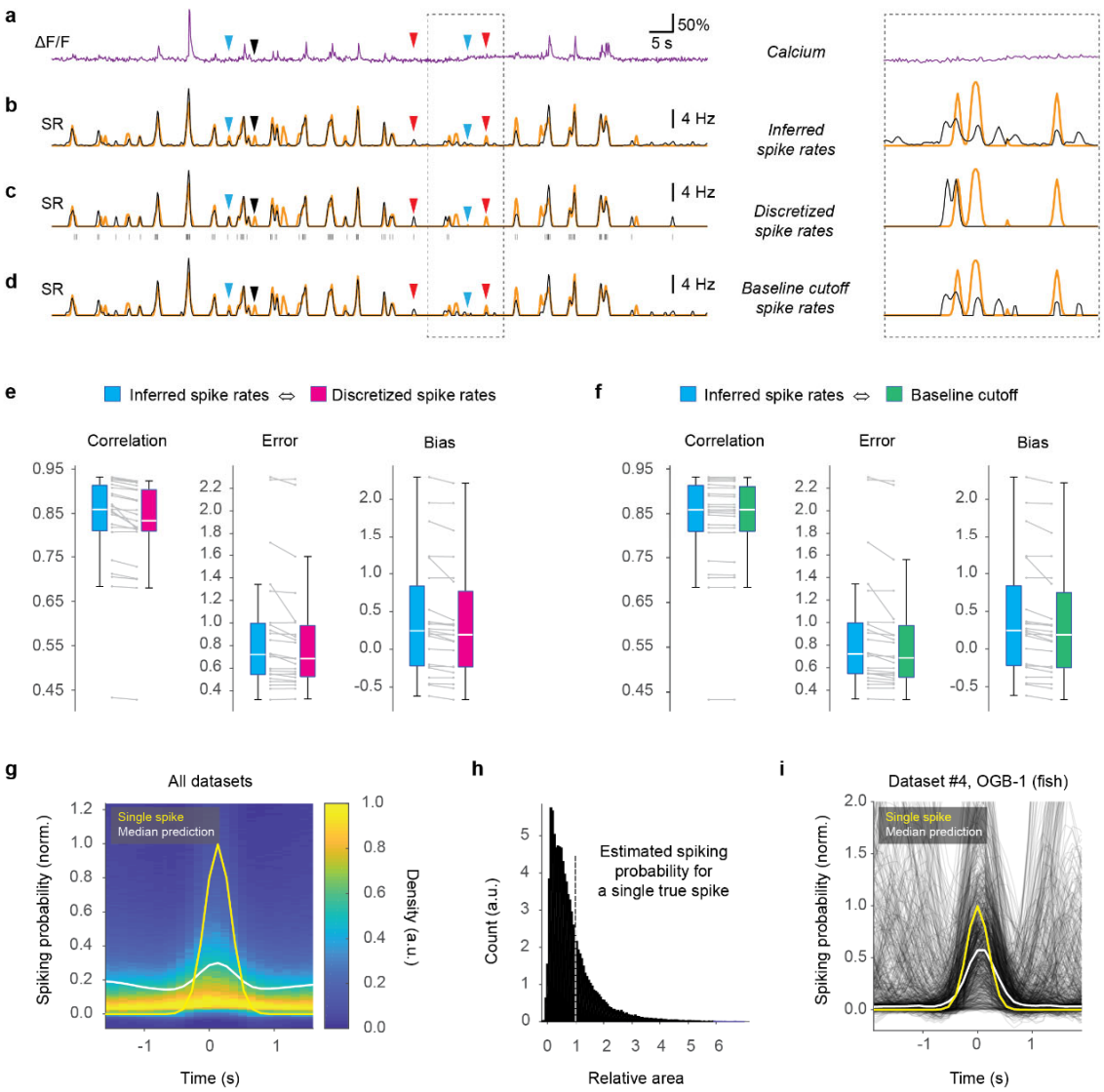

Figure S7 | Analysis of single-spike resolution of spike inference. a, Calcium $\Delta F/F_0$ of an example neuron. **b,** Inferred spike rate (SR, black) and ground truth spike rate (orange). Inset to the right zooms into a section of the recording to highlight the noise floor. **c,** Discrete spikes were fitted into the inferred spiking probabilities from (b). Blue arrowheads indicate predictions that were therefore improved, red arrowheads indicate predictions that there degraded. **d,** Spiking probabilities from (b), but using a threshold (1/e of the magnitude of a single spike) as cutoff for the noise floor. **e,** Comparison of inferred probabilities (cf. (b)) vs. discretized spiking (cf. (c)). While errors and biases are slightly decreased for discretized spiking, also the correlations are reduced. Each data point is the median across a ground truth dataset, not including the datasets with inhibitory neurons (n=22 datasets). **f,** Thresholding the predictions as in (d) results in the same reduction of errors and biases, without negatively affecting the correlations (n=22 datasets). For box plots in (e) and (f), the median is indicated by the central line, 25th and 75th percentiles by the box, and maximum/minimum values excluding outliers (points) by the whiskers. **g,** Distribution of predicted spiking probability for an isolated action potential in the ground truth across all datasets. Across all datasets, the estimate (white) of a single ground truth action potential is clearly lower than expected from ground truth (yellow). **h,** Distribution of the overall area under the curve associated with a single isolated ground truth action potential. In an ideal case, values would be narrowly distributed around 1 (dashed line). Instead, the distribution is broad and biased with a central value <1. **i,** Spiking probability for a single dataset that should be ideally qualified for single-spike precision (DS#04; no movement artifacts due to *ex vivo* preparation, synthetic indicator OGB-1). Even here, the median prediction is on average systematically lower than a single action potential, indicating that the network trades off single-spike precision in order to prevent false positive detection of spikes amidst of shot noise.

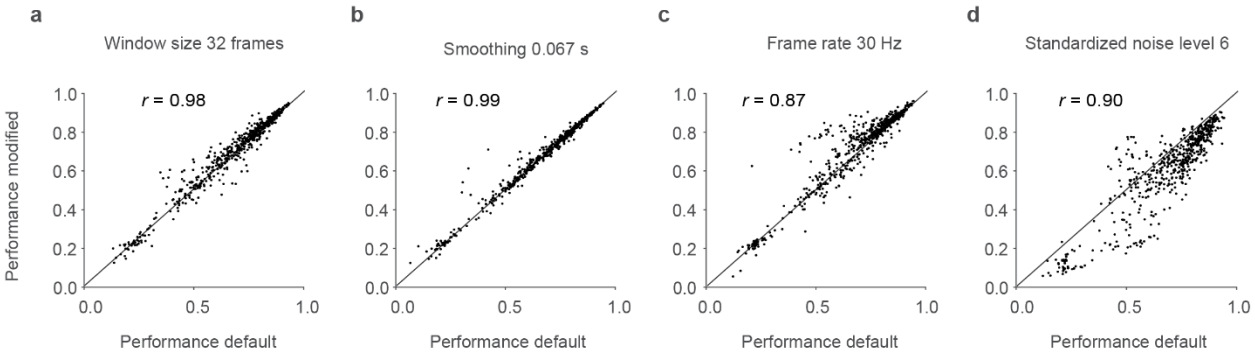

Figure S8 | Parameter changes leave mutual predictability across datasets largely unchanged. The off-diagonal elements in Fig. 4a were recomputed using different parameter settings. The standard parameters were a smoothing kernel of 0.2 s, a framerate of 7.5 Hz, a standardized noise level of 2 and a window size of the deep network of 64 data points. The performance (correlation with ground truth) for the standard parameter analysis ("default", x-axis) was plotted against the performance for modified parameters. Each data point corresponds to an off-diagonal matrix element in Fig. 3a. The red line indicates the unity function. Correlation between the two conditions is indicated as *r*. **a,** Window size decreased to 32. **b,** Smoothing kernel reduced to 0.067 s. **c,** Frame rate increased to 30 Hz. **d,** Standardized noise level increased to 6.

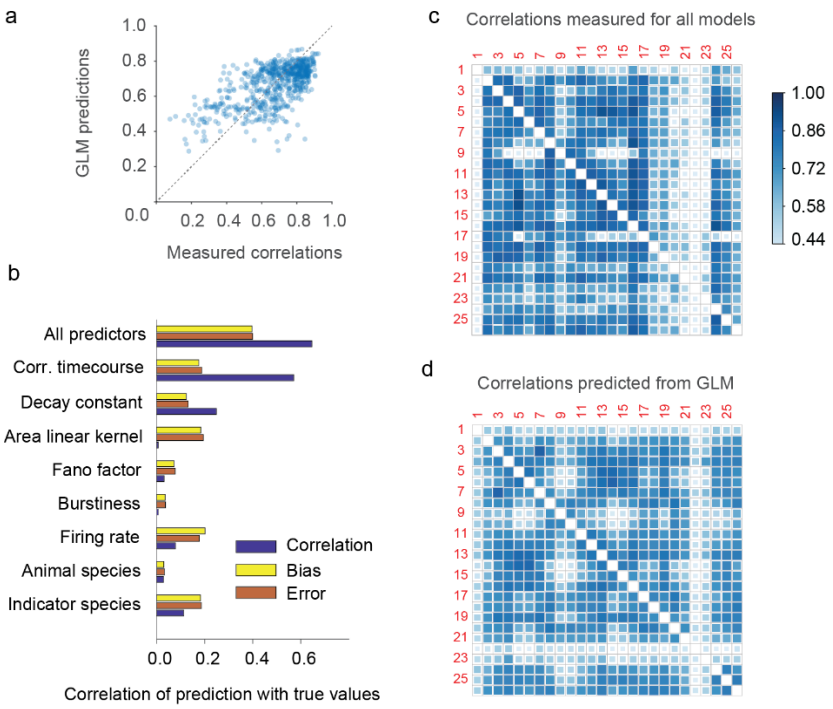

Figure S9 | Predicting cross-dataset predictability with a generalized linear model (GLM). We considered several characteristics of ground truth datasets and evaluated whether they could improve the mutual predictability across ground truth datasets. This includes characteristics that are accessible without ground truth (indicator species, i.e., synthetic dyes vs. genetically encoded indicators; animal species, *i.e.,* mouse vs. zebrafish; median spike rate across neurons; the burstinesss; the Fano factor) and characteristics that are only accessible with available ground truth (area under the curve of the linear kernel, cf. Extended Data Fig. 1; decay constant of the linear kernel; the correlation between the kernels of the training and test datasets), with detailed descriptions in the Methods section. These 8 predictors were used as regressors for a GLM to fit the mutual predictability matrix (correlations) among datasets. **a,** Correlations predicted from the GLM vs. measured correlations (see Fig. 3a). **b,** A GLM based on "all predictors" results in a correlation that recovers panel (a). In addition, this plot also shows values for fits to the bias and error matrices. Using only one of the predictors reduces the correlation, often very significantly. **c,** Measured cross-dataset predictability (reproducing Fig. 3a). **d,** Cross-dataset predictability as computed with the GLM. Together, this supplementary figure shows that the GLM is not able to explain a large fraction of the variance of the original matrix, and that the main predictor is the correlation between kernel time courses, which is not accessible without available ground truth. As consequence, we chose the use of a "global EXC model" that was trained on all reliable datasets (Fig. 3a).

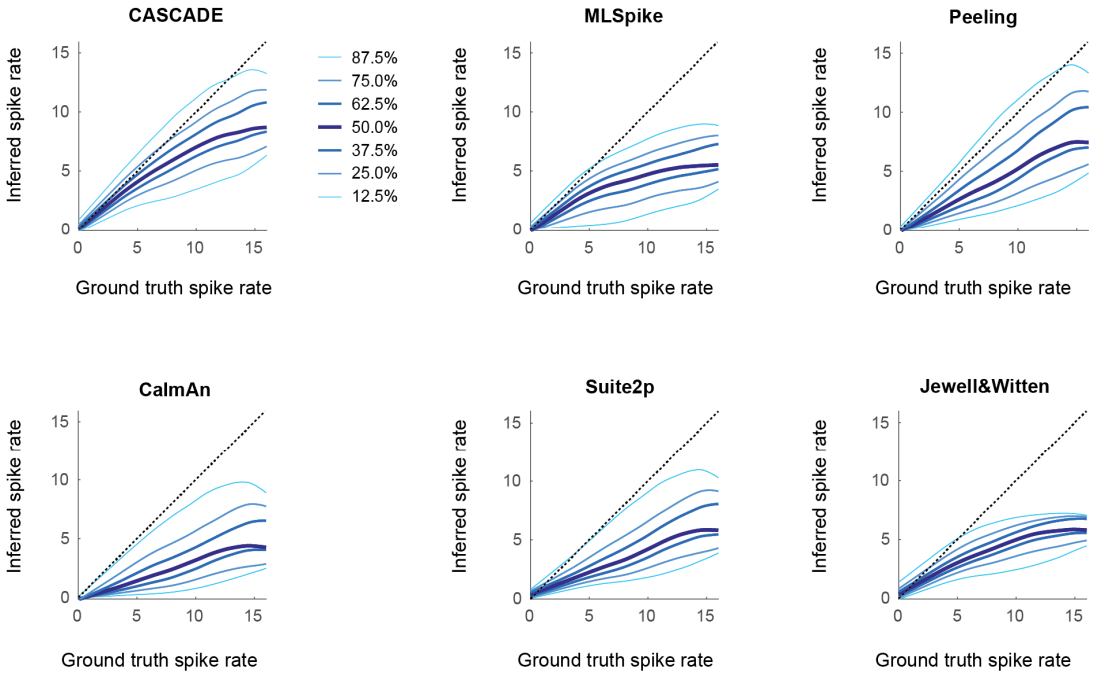

Figure S10 | Bias of predictions across spike rates. Spike rates averaged for per 2 s-window, ground truth spike rate vs. Inferred spike rates for all algorithms. For each bin on the x-axis, a histogram was generated. Lines indicate the percentiles of the respective histograms. The 50th percentile recovers the medians shown in Fig. 4g.

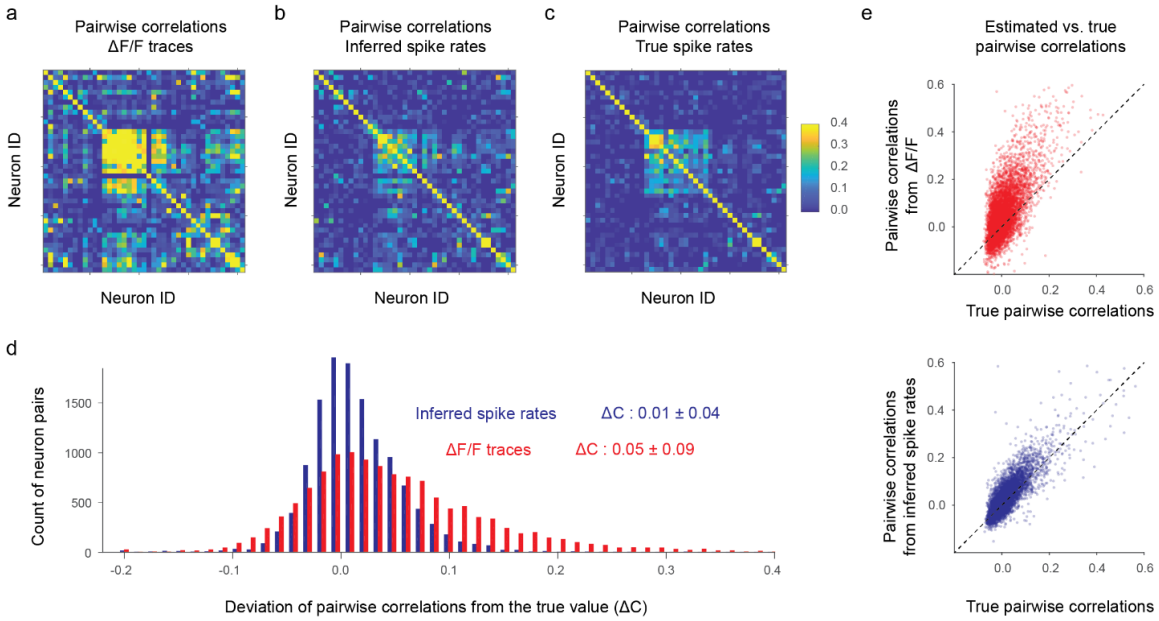

Figure S11 | Spike inference with CASCADE improves pairwise correlations for population imaging. To analyze how spike inference from calcium imaging data affects population analyses and, more specifically, pairwise correlations between neurons, we used NAOMi to generate artificial calcium imaging data that includes neuropil activity, realistic photon detection schemes and indicator kinetics for GCaMP6f (see Methods) at a frame rate of 30 Hz. Spike rates were inferred by the "Global EXC model" (Fig. 3) which excluded the NAOMi ground truth dataset. **a**, Pairwise correlations between neurons in a simulated imaging experiment with 43 neurons, computed from the ΔF/F traces. **b**, Pairwise correlations, computed from the inferred spike rates. **c**, Pairwise correlations, computed from the true spike rates. **d**, Deviation (ΔC) of pairwise correlations from the true value for ΔF/F traces (red) and inferred spike rates (blue). Pairwise correlations derived from ΔF/F traces were biased towards higher values (mean deviation and standard deviation of ΔC: 0.05 ± 0.09). However, also underestimates of pairwise correlations were more frequent than for pairwise correlations derived from inferred spike rates. Pairwise correlations estimated from inferred spike rates were in general closer to the true values and less biased towards overestimates (0.01 ± 0.04). **e**, Pairwise correlations estimated from ΔF/F traces (red; top panel) and inferred spike rates (blue; bottom panel) plotted against the true pairwise correlation for each neuron pair. Estimates from inferred spike rates were closer to true pairwise correlations (unity line, dashed). Panels (a-c) are based on a single imaging volume simulated with NAOMi. Panels (d-e) are based on five simulated imaging volumes with a total of 250 extracted neurons (see Methods), resulting in a total of 17,506 neuron pairs.

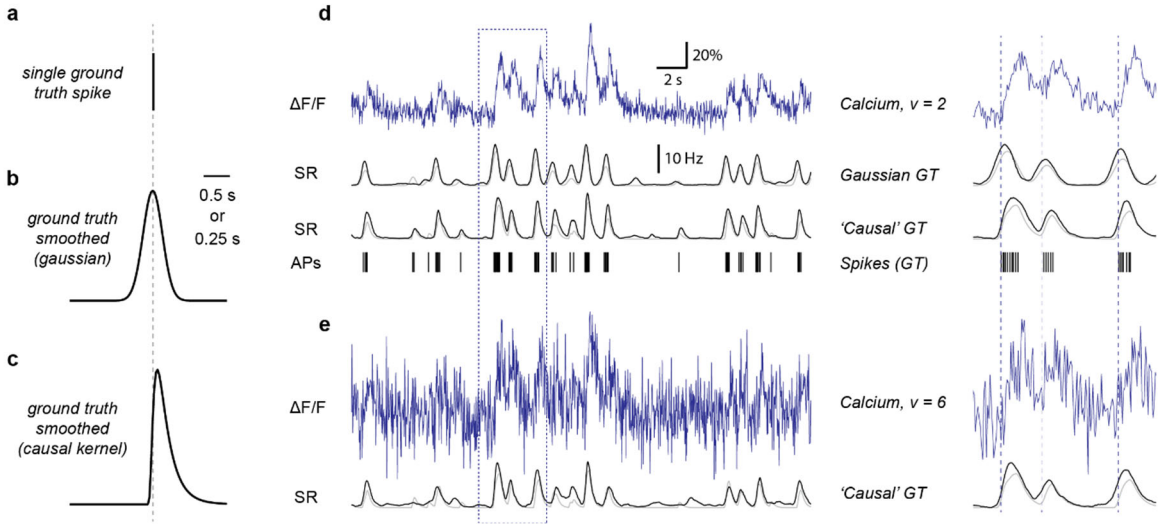

Figure S12 | Non-Gaussian ground truth smoothing kernel. Since calcium signals rise after action potentials, the task for spike inference is to re-assign ΔF/F activity to the true spike time. Due to noise and imperfect match between model and calcium trace, the re-assignment is typically slightly off, into both the past and the future of the true spike. In principle, both re-assignment to past and future is equally undesirable. In some scenarios, however, re-assignment of the activity into the past of the spike can be particularly adverse, for example, when a precise external stimulus (*e.g.*, an auditory tone) is used to generate a peri-stimulus activity trace. While the ground truth used for training **(a)** is typically smoothed symmetrically in time with a Gaussian function **(b)**, by design activity that happened briefly after the stimulus will be deconvolved to time points both after and before the stimulus. To circumvent these a-causal events, it is possible to use a more causal filter **(c)**, using a highly skewed inverse Gaussian distribution. **d,** The deep network can use the 'causal' ground truth, resulting in inferred spike rates that re-assign activity in a more causal way (lower prediction trace; see zoom-in). However, noise sources result in re-assignment of activity to time bins prior to ground truth spikes. **e,** The predictions become necessarily more sloppy and smeared out for higher noise levels. This shows that even when using a causal kernel to smooth the ground truth, a-causal effects can be induced by spike inference. Depending on the desired temporal resolution of the algorithm, different width of the smoothing kernels can be chosen. A Gaussian with $\sigma = 0.2$ s will lead to a FWHM of the smoothing Gaussian of ca. 0.5 s (b). For panel (d), a smoothing kernel with $\sigma = 0.2$ s was used.

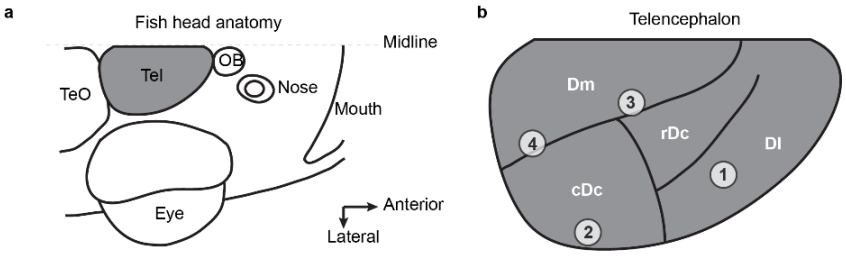

Figure S13 | Locations of neurons in the dorsal telencephalon of adult zebrafish from ground truth dataset DS#08. a, Fish head anatomy, highlighting the Telencephalon from a dorsal aspect. TeO = optic tectum, OB = olfactory bulb, Tel = Telencephalon. **b,** Recording locations in the telencephalon (see Methods for details). Neuron #1 of the ground truth dataset was located in Dl, ca. 120 µm below the dorsal surface of Dm (position 1). Neurons #2-#5 were located in cDc, ca. 50 µm below the dorsal surface of Dm (position 2). Neurons #6 and #7 were located close to the sulcus ypsilonformis at the interface of Dm and rDc, at a depth of ca. 80 µm below the dorsal surface of Dm (location 3). Neurons #8, #9 and #10 were located at the interface of Dm and cDc, ca. 20-30 µm below the dorsal surface of Dm. Nomenclature and subdivisions follow Huang et al. (2020).

NOISE LEVEL 2	MLSpike		Peeling		CalmAn	Suite2p		Jewell	
	tau	amplitude	tau	amplitude	gamma	sig_baseline	gamma	penalty	gamma
DS01	0.1000	0.3500	0.6360	7.2730	0.0200	0.2500	0.0250	0.0040	0.4730
DS02	1.9690	0.3500	0.2500	31.5620	0.0200	32.5000	0.2190	0.0200	0.7060
DS04	1.5000	0.1910	1.0170	16.0000	0.0250	20.6670	0.1000	0.0500	0.7000
DS05	4.0000	0.0100	2.3750	35.0000	0.7400	33.7500	0.6750	0.0840	0.8500
DS06	2.0000	0.0900	2.3750	31.8750	0.8030	0.2500	0.6750	0.0070	0.9500
DS07	2.0000	0.2420	2.1000	35.0000	0.8180	0.2500	0.6900	0.0160	0.9200
DS08	1.2220	0.0300	2.0000	8.3330	0.1890	0.2500	0.0920	0.0040	0.6560
DS09	1.0000	0.2500	0.9320	28.6360	0.4180	4.3180	0.2000	0.0500	0.5090
DS10	0.7500	0.1610	0.8040	35.0000	0.5540	0.2500	0.2830	0.0100	0.8000
DS11	0.7600	0.1300	0.5300	29.4000	0.1130	0.3100	0.1000	0.0150	0.7180
DS12	3.6670	0.3500	2.4170	35.0000	0.7730	2.1250	0.5500	0.0400	0.8000
DS13	2.0000	0.3500	2.9810	35.0000	0.8170	0.2500	0.7920	0.0200	0.9000
DS14	1.0000	0.2500	0.9320	28.6360	0.4180	4.3180	0.2000	0.0500	0.5090
DS15	1.5000	0.0300	2.4440	15.5560	0.6220	0.2500	0.3000	0.0070	0.9000
DS16	0.7500	0.0100	1.1110	14.4440	0.0420	0.2500	0.0500	0.0040	0.7110
DS17	0.8330	0.0830	1.5000	20.0000	0.5040	0.3610	0.2000	0.0100	0.8000
DS18	1.0630	0.0700	1.1250	18.7500	0.3200	0.2500	0.1810	0.0040	0.9130
DS19	1.2500	0.1060	0.5280	18.8890	0.0200	2.5000	0.0250	0.0190	0.5890
DS20	2.5000	0.0860	1.9440	10.5560	0.1620	4.7220	0.1220	0.0320	0.7890
DS21	0.4050	0.2460	0.5000	35.0000	0.0200	4.0910	0.0390	0.0550	0.2320
DS22	1.6200	0.3500	0.4000	10.0000	0.1880	4.6000	0.0250	0.0060	0.2400
DS23	2.1570	0.1790	0.2860	2.5000	0.0200	0.2500	0.0250	0.0010	0.1000
DS24	0.2000	0.0130	2.0390	2.5000	0.0200	0.2500	0.0250	0.0030	0.1810
DS25	0.2000	0.0290	1.8530	8.8240	0.0690	0.2500	0.0290	0.0030	0.5650
DS26	0.2000	0.0550	1.1250	9.3750	0.0200	0.2500	0.0250	0.0040	0.4000

NOISE LEVEL 9	MLSpike		Peeling		CalmAn	Suite2p		Jewell	
	tau	amplitude	tau	amplitude	gamma	sig_baseline	gamma	penalty	gamma
DS01	0.1000	0.3500	0.3640	19.0910	0.2450	5.4550	0.0250	0.0250	0.1000
DS02	0.1130	0.3500	4.9380	35.0000	0.7600	50.0000	2.8940	0.2500	0.9000
DS04	2.0000	0.3500	0.5000	29.3330	0.0200	30.0000	0.5130	0.2500	0.3500
DS05	2.6250	0.2650	2.2500	35.0000	0.7150	45.0000	1.2250	0.5630	0.7250
DS06	1.9380	0.0300	2.0000	34.3750	0.7830	4.3750	0.8000	0.1190	0.8000
DS07	2.9000	0.1580	2.5000	35.0000	0.8160	3.0000	1.0300	0.2500	0.8000
DS08	1.7220	0.0830	2.2780	8.3330	0.1760	0.3610	0.1830	0.0300	0.1170
DS09	1.4770	0.3500	0.8860	33.1820	0.5380	50.0000	0.6450	0.2500	0.5000
DS10	0.7500	0.1600	1.0000	35.0000	0.5530	3.4780	0.3740	0.1000	0.2480
DS11	0.5000	0.2880	1.0000	20.0000	0.1460	5.2000	0.2000	0.1000	0.1900
DS12	1.2920	0.3500	2.4170	35.0000	0.7770	3.8330	1.1330	0.2250	0.7000
DS13	1.5000	0.2270	3.0000	35.0000	0.8450	0.9620	1.3270	0.1810	0.7460
DS14	1.4770	0.3500	0.8860	33.1820	0.5380	50.0000	0.6450	0.2500	0.5000
DS15	1.5000	0.0300	2.3890	15.5560	0.5910	1.5000	0.3890	0.1000	0.6670
DS16	1.5000	0.0300	1.0000	13.0560	0.0200	0.3060	0.1110	0.0250	0.1610
DS17	0.9720	0.0720	1.7220	19.4440	0.5000	17.2220	0.2560	0.1000	0.5000
DS18	1.0630	0.3200	1.6250	14.3750	0.3200	1.1250	0.2880	0.0590	0.4630
DS19	1.8890	0.3480	0.5000	18.8890	0.0200	18.5710	0.2140	0.1000	0.2640
DS20	2.5000	0.2260	2.0560	15.0000	0.1130	18.8890	0.6000	0.2500	0.5670
DS21	0.4320	0.3010	0.5000	33.1820	0.0200	29.0910	0.0950	0.1140	0.1000
DS22	0.1000	0.3500	3.2500	33.0000	0.3200	34.0000	1.5400	0.1500	0.5100
DS23	0.1000	0.3500	2.4290	3.5710	0.1290	0.2500	0.0290	0.0000	0.2290
DS24	0.3500	0.0300	2.0770	2.5000	0.0200	0.2500	0.0250	0.0030	0.1000
DS25	0.3590	0.0310	1.5880	9.4120	0.0550	0.6320	0.0570	0.0200	0.1000
DS26	0.1000	0.0700	1.0000	10.0000	0.0200	0.5630	0.0250	0.0100	0.1000

Supplementary Table 1 | Optimal parameters for the model-based spike detection algorithms. Values were found for each dataset separately using a grid search over the indicated parameters (Methods) for standardized noise levels $\nu = 2$ (top) and $\nu = 9$ (bottom). Standardized nose levels are measured in units of $\% \cdot Hz^{-1/2}$. Columns contain optimal parameters obtained by minimizing the mean squared error of predictions with respect to ground truth (frame rate 7.5 Hz, smoothing with σ = 0.2 s).