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

Peter Rupprecht [1,2,†], Stefano Carta [1], Adrian Hoffmann [1], Mayumi Echizen [4,5], Kazuo Kitamura [4,6], Fritjof Helmchen [1,†,*], Rainer W. Friedrich [2,3,†,*]

[1] Brain Research Institute, University of Zürich, Zürich, Switzerland
[2] Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland
[3] University of Basel, Basel, Switzerland
[4] Department of Neurophysiology, University of Tokyo, Tokyo, Japan
[5] Department of Anesthesiology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan
[6] Department of Neurophysiology, University of Yamanashi, Yamanashi, Japan
* Contributed equally

† Corresponding authors: rupprecht@hifo.uzh.ch, helmchen@hifo.uzh.ch, rainer.friedrich@fmi.ch

ABSTRACT

Calcium imaging is a key method to record patterns of neuronal activity across populations of identified neurons. Inference of temporal patterns of action potentials ('spikes') from calcium signals is, however, challenging and often limited by the scarcity of ground truth data containing simultaneous measurements of action potentials and calcium signals. To overcome this problem, we compiled a large and diverse ground truth database from publicly available and newly performed recordings. This database covers various types of calcium indicators, cell types, and signal-to-noise ratios and comprises a total of >20 hours from 225 neurons. We then developed a novel algorithm for spike inference (CASCADE) that is based on supervised deep networks, takes advantage of the ground truth database, infers absolute spike rates, and outperforms existing model-based algorithms. To optimize performance for unseen imaging data, CASCADE retrains itself by resampling ground truth data to match the respective sampling rate and noise level. As a consequence, no parameters need to be adjusted by the user. To facilitate routine application of CASCADE we developed systematic performance assessments for unseen data, we openly release all resources, and we provide a user-friendly cloud-based implementation.

INTRODUCTION

Imaging of somatic calcium signals using organic or genetically encoded fluorescent indicators has emerged as a key method to measure the activity of many identified neurons simultaneously in the living brain[1,2]. However, calcium signals are only an indirect, often non-linear and low pass-filtered proxy of the more fundamental variable of interest, somatic action potentials (spikes) [3–5]. The relationship between calcium signals and spike rates is ideally assessed directly by

simultaneous electrophysiological recordings, preferably in the minimally disruptive juxtacellular configuration, and optical imaging of a calcium indicator signal in the same neuron. These dual recordings can serve as ground truth to calibrate and optimize algorithms for the inference of spike rates from other calcium imaging data (Fig. 1a). Based on such ground truth datasets, various model-based methods[6–16,12] as well as supervised machine learning algorithms[15,17–19] for spike inference have been developed.

Ideally, an algorithm should be applicable to infer spike rates in unseen calcium imaging datasets for which no ground truth is available. The relationship between spikes and the evoked calcium signals depends on multiple factors including the neuron type, the type of calcium indicator and its concentration, the optical resolution, the sampling rate and the noise level. Many of these parameters can vary substantially between experiments and even from neuron to neuron within the same experiment. As a consequence, experimental conditions of novel datasets are often not well matched to those of available ground truth data. It is therefore not clear how an algorithm based on a specific ground truth dataset generalizes to other datasets, which causes problems for the inference of spike rates from calcium imaging data under most experimental conditions[12,13,20,21].

Here, we address the issue of generalization systematically. To assemble a large ground truth database, we performed juxtacellular recordings and two-photon calcium imaging using different calcium indicators and in different brain regions of zebrafish and mice. This database was then augmented with a carefully curated selection of publicly available ground truth datasets. Using this large database, we developed a supervised method for **ca**librated **s**pike inference of **ca**lcium data using **de**ep networks (CASCADE). CASCADE includes methods to resample the original ground truth datasets in order to match their sampling rate and noise level to a specific calcium imaging dataset of interest. This procedure allowed us to train machine learning algorithms upon demand on a broad spectrum of resampled ground truth datasets, matching a wide range of experimental conditions. Finally, we tested the performance of CASCADE systematically when applied to unseen data. CASCADE was robust with respect to any hyper-parameter choices and performed better than existing algorithms in a benchmark test performed across all ground truth datasets. The CASCADE algorithm can be used directly via a cloud-based web application and is also available together with the ground truth datasets as a simple and user-friendly Python-based toolbox.

RESULTS

A large dataset of curated ground truth recordings

To extend the spectrum of existing ground truth datasets, we performed simultaneous electrophysiological recordings and calcium imaging in adult zebrafish and mice (Fig. 1b-h; Table 1). In zebrafish, a total of 47 neurons in different telencephalic regions were recorded in the juxtacellular configuration in an explant preparation of the whole adult brain[22] using the synthetic

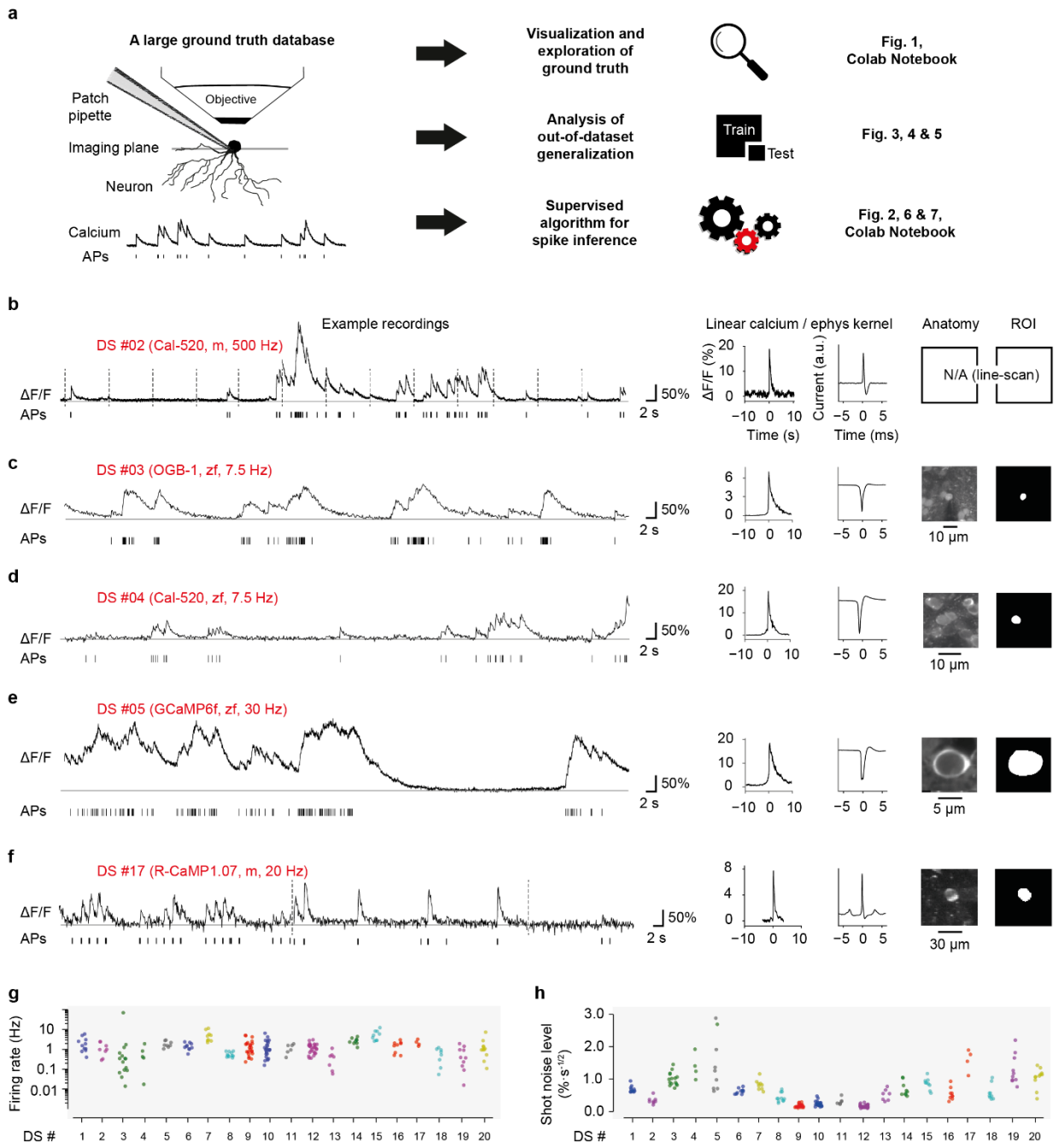

Figure 1 | **Ground truth datasets. a,** A large and diverse ground truth database obtained by simultaneous calcium imaging and juxtacellular recording (left) can be used for 1) the exploration of the ground truth by a user, for 2) the analysis of the out-of-dataset generalization of spike inference and for 3) the training of a supervised algorithm for spike inference. The right column refers to relevant figures. *Colab Notebook* refers to relevant cloud-based tools accompanying this paper. **b-f,** Examples of ground truth recordings with different indicators, different brain regions and species. Left: calcium signal traces (ΔF/F) are shown together with the detected action potentials (APs). Middle: linear kernels of ΔF/F (time scale in seconds) and electrophysiological data (time scale in milliseconds) triggered by single spikes. Right: fluorescence image of the respective neuron, together with the ROI for fluorescence extraction. **g,** Average spike rate for each neuron of the ground truth database (log scale). 20 datasets (DS) were included in total. **h,** Shot noise level for each neuron of the ground truth. Shot noise levels were computed as the median absolute difference between subsequent time points of a ΔF/F trace.

Dataset identifier	Calcium indicator	Induction method	Animal model	Brain region	Frame rate [Hz]	High-freq noise [%·s$^{-1/2}$]	Spike rate [Hz]	# of neurons	Recording duration [min]	Source paper
#1	OGB-1	acute injection	Mouse	V1	11.3	0.7±0.2	5.5±1.5	11	83	Theis et al., 2016
#2	Cal-520	acute injection	Mouse	S1	500.0	0.3±0.1	1.2±0.8	8	23	this paper; Tada et al., 2014
#3	OGB-1	acute injection	Zebrafish	pDp	7.7	1.0±0.2	0.4±0.5	15	81	this paper
#4	Cal-520	acute injection	Zebrafish	pDp	7.8	2.0±1.3	1.3±2.1	5	31	this paper
#5	GCaMP6f	tg(NeuroD)	Zebrafish	aDp	30.0	1.3±0.8	1.9±0.7	8	46	this paper
#6	GCaMP6f	tg(NeuroD)	Zebrafish	dD	30.0	0.6±0.1	1.5±0.6	10	69	this paper
#7	GCaMP6f	tg(NeuroD)	Zebrafish	OB	30.0	0.8±0.2	5.3±3.3	9	45	this paper
#8	GCaMP6f	AAV	Mouse	V1	60.1	0.4±0.1	0.6±0.2	11	129	Chen et al., 2013
#9	GCaMP6f	tg(Emx1)	Mouse	V1	160.1	0.5±0.2	1.6±1.4	23	72	Huang et al., 2019
#10	GCaMP6f	tg(Cux2)	Mouse	V1	158.3	0.5±0.2	1.5±1.5	25	78	Huang et al., 2019
#11	GCaMP6s	tg(tetOs)	Mouse	V1	151.6	0.8±0.1	1.0±0.4	6	13	Huang et al., 2019
#12	GCaMP6s	tg(Emx1)	Mouse	V1	157.5	0.5±0.2	1.3±0.7	26	62	Huang et al., 2019
#13	GCaMP6s	AAV	Mouse	V1	60.1	0.5±0.2	0.4±0.4	7	70	Chen et al., 2013
#14	GCaMP6s	AAV	Mouse	V1	59.1	0.7±0.2	6.2±3.5	9	77	Theis et al., 2016
#15	GCaMP6s	AAV	Mouse	V1	59.1	0.9±0.2	5.8±3.3	9	25	Theis et al., 2016
#16	GCaMP5k	AAV	Mouse	V1	50.0	0.5±0.2	1.6±0.9	9	29	Akerboom et al., 2012
#17	R-CaMP1.07	tg(Grik4-cre) AAV virus	Mouse	CA3	20.0	1.6±0.3	2.2±0.8	4	33	this paper
#18	R-CaMP1.07	AAV	Mouse	S1	15.0	0.6±0.2	0.9±1.0	9	50	this paper; Bethge et al., 2017
#19	jRCaMP1a	AAV	Mouse	V1	15.0	1.3±0.5	0.6±0.6	10	88	Dana et al., 2016
#20	jRGECO1a	AAV	Mouse	V1	29.8	1.0±0.3	1.6±2.0	11	118	Dana et al., 2016

Table 1 | Overview of all ground truth datasets. High-frequency noise was determined as described in Methods. Frame rate was given as the mean across experiments if the frame rates varied (typically only slightly) across experiments within a single dataset. Noise levels and spike rates are given as mean ± s.d. across neurons.

calcium indicators Oregon Green BAPTA-1 (OGB-1) and Cal-520 as well as the genetically encoded calcium indicator GCaMP6f. We also performed ground truth recordings in head-fixed, anesthetized mice in hippocampal area CA3 using the genetically encoded indicator R-CaMP1.07, and we extracted ground truth recordings from raw data of previous publications based on Cal-520 and R-CaMP1.07, respectively, in mouse primary somatosensory cortex (S1; total of 21 neurons) [23,24]. In addition, we surveyed openly accessible datasets and extracted ground truth from raw movies (when available) or preprocessed calcium imaging data[15,18,25–29]. Rigorous quality control (Methods) reduced the original number from a total of 193 available

neurons to 157 neurons. Together with our own recordings, we have assembled 20 datasets with a total of 225 neurons, spanning 8 calcium indicators and 9 brain regions in 2 species, totaling >20 hours of recording and 146,724 spikes.

Recording durations, imaging frame rates and spike rates varied greatly across the ground truth datasets (Table 1). Typical spike rates spanned an order of magnitude, ranging from 0.4 to 6.2 Hz, and frame rates varied between 7.7 Hz and >160 Hz (Table 1; Fig. 1g). We used regularized deconvolution to compute the linear ΔF/F kernel evoked by the average spike and found that the area under the kernel curve varied substantially across datasets, even when data were obtained with the same indicator. In fact, kernels varied substantially even across neurons within the same dataset (Fig. S1). This diversity highlights the challenge faced by any algorithm that is supposed to generalize to unseen data.

Inference of spike rates with a deep convolutional network

Several favorable properties make supervised deep learning approaches well suited for spike inference from calcium imaging data. First, deep learning generally tends to outperform other classification or regression methods if the amount of training data is sufficiently high (typically >1000 data points for each category in classification tasks) [30–32]. Second, the cost function can easily be modified to optimize the metric of interest, *e.g.*, correlation with ground truth or mean squared error, without changing network architecture. Third, the temporal extent of receptive fields of deep networks can be adapted to account for history-dependent effects such as the dependence of action potential-evoked calcium transients on previous activity (see Fig. S2 for an example). Finally, deep networks are intrinsically non-linear, allowing to fit non-linear behaviors of calcium indicators.

We designed a simple convolutional network that uses a segment of the calcium signal trace (expressed as percentage fluorescence change ΔF/F) around a time point t to infer the probability that a spike has occurred at t. Compared to two-dimensional image classification and object labeling[30,33,34], requirements on computational hardware are low because datasets are small and the inference task is only one-dimensional (time). For example, ImageNet[35], a dataset used for visual object identification and detection in the deep learning field, is typically used at a resolution of 256 x 256 = 65,536 data points per sample, whereas the input used for spike inference in this study was smaller by approximately three orders of magnitude, typically consisting of a segment of the ΔF/F trace with 64 data points.

We used a network architecture with a standard convolutional design, consisting of rectifying linear units (ReLUs) that were distributed across three convolutional layers, two pooling layers and a single dense layer. The final dense layer projected to a single output unit that reported the estimated spike rate for the current time t (Fig. 2a; see Methods for more details).

Resampling of ground truth data for noise-matching

The key idea underlying our approach is that the ground truth (training data) is as important as the algorithm itself and should match as well as possible the noise level and sampling rate of the unseen population calcium data of interest (test data). We therefore devised a workflow where noise level and sampling rate are extracted from the test data and then used to generate noise- and rate-matched training data from the ground truth database (Fig. 2b). To facilitate gradient descent, the ground truth spike rate is smoothed with a Gaussian kernel (Methods).

To extract ΔF/F noise levels, we computed a standardized noise metric ν that is robust against outliers and approximates the standard deviation of ΔF/F baseline fluctuations. To allow for comparison of noise measurements across datasets, ν was normalized by the square root of the sampling rate, resulting in a natural unit of $[\nu] = \% \cdot s^{-1/2}$ (for details see Methods; Fig. S3 for illustration of typical noise levels). To generate training data with pre-defined ΔF/F noise levels, we explored several approaches based on sub-sampling of ROIs or additive artificial noise (Supplementary Note 1; Fig. S4). We identified the addition of artificial Poisson-distributed noise as the most suitable approach to transform the ground truth data into appropriate training data for the deep network.

To quantify deep network performance, we developed a set of complementary metrics for the accuracy of spike inference. Following previous studies, we calculated the Pearson **correlation** between ground truth spike rates and inferred spike rates[15,18]. However, this measure leaves the absolute magnitude of the inferred instantaneous spike rate unconstrained. We therefore also computed the positive and negative deviations of the absolute spike rate from the ground truth spike rate. The sum of the absolute deviations was defined as the **error** while the sum of the signed deviations was defined as the **bias** of the inference (see Methods and Fig. S5). Error and bias were both normalized by the number of true spikes to obtain relative metrics that can be compared between datasets. While the correlation is arguably the most important metric because it estimates the similarity of inferred and true spike rates, error and bias are important to assess the inference of absolute spike rates because they identify spike rates that are either incorrectly scaled or systematically too large or small.

We found that the performance of the deep network degraded considerably when the noise level of the test dataset deviated considerably from the noise level of the ground truth. As expected by intuition, a network that had only seen almost noise-free data during training failed to suppress fluctuations of noisier recordings. Conversely, we also observed that a network trained on very noisy calcium signals was unable to fully benefit from low-noise calcium recordings, inferring only an imprecise approximation of the ground truth (Fig. 2c). A systematic iteration across combinations of noise levels for training and test datasets showed that for each test noise level the best model had been trained with a similar or slightly higher noise level (Fig. 2d-g; even more apparent when normalizing the metrics for each test noise level in Fig. S6). Very low noise levels ($\nu < 2$) result in a special case (Fig. 2d,e): since some neurons of a given ground truth dataset do not reach the desired noise level even without addition of noise (cf. Fig. 1h), the effective size of the training dataset decreases, resulting in slightly lower performance. In general, however, the

results show that it is beneficial to train with noise levels that are adapted to the calcium data for which the algorithm will be applied after training.

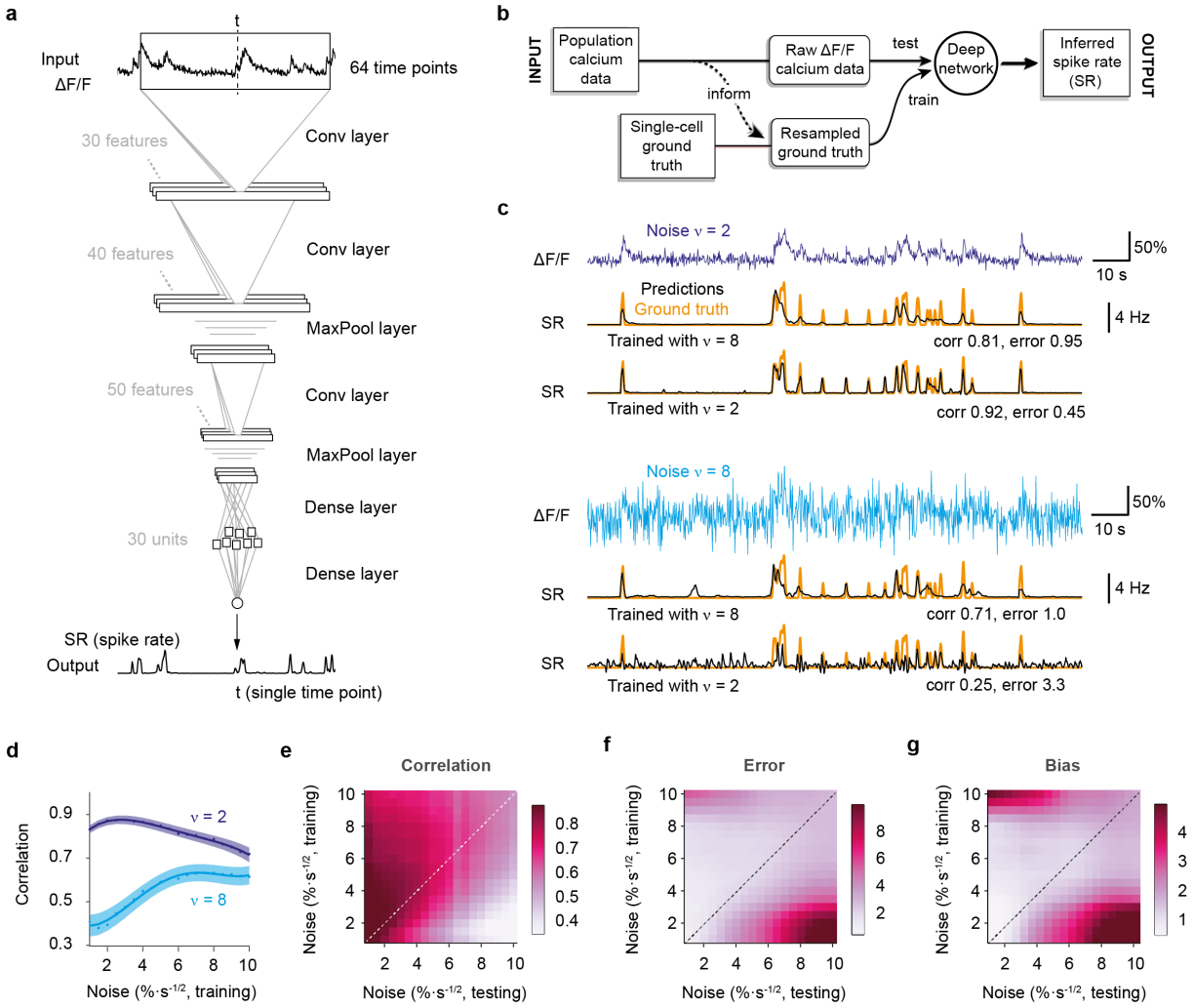

Figure 2 | Training a deep network with noise-matched ground truth improves spike inference. a, The default deep network consists of an input time window of 64 time points centered around the time point of interest. Through three convolutional layers, two pooling layers and one small dense layer, the spiking probability is extracted from the input time window and returned as a single number for each time point. **b,** Properties of the population data (frame rate, noise level; dashed line) are extracted and used for noise-matched resampling of existing ground truth datasets. The resampled ground truth is used to train the algorithm, resulting in calibrated spike inference of the population imaging data. **c**, Top: a low-noise ΔF/F trace is translated into spike rates (SR; inferred spike rates in black, ground truth in orange) more precisely when low-noise ground truth has been used for training. Bottom: a high-noise ΔF/F trace is translated into spike rates (SR; inferred spike rates in black, ground truth in orange) more precisely when high-noise ground truth has been used for training. **d,** The spike inference performance for two test conditions (low noise, $\nu = 2$, dark blue; high noise, $\nu = 8$, light blue) is optimal when training noise approximates testing noise levels. **e,** Correlation between predictions and ground truth is maximized if noise levels of training datasets match noise levels of testing sets. See also Fig. S6. **f,** Relative error of predictions with respect to ground truth. **g,** Relative bias of predictions with respect to ground truth.

Parameter-robustness of spiking inference

Traditional models to infer spiking activity typically contain a small number of parameters[10–12,14] that describe biophysical quantities and are adjusted by the user. Deep networks, in contrast, contain thousands or millions of parameters adjusted during training that have no obvious biophysical meanings[13,15]. The user can modify only a small number of hyper-parameters that define general properties of the network such as the loss function, the number of features per layer, or the receptive field size. We therefore tested how spike inference performance depends on these hyper-parameters.

We found that the performance of the network was robust against variations of all hyper-parameters (Supplementary Note 2; Fig. S7a-e). Moreover, overfitting was moderate despite prolonged training, indicating that the abundance of noise and sparseness of events act as a natural regularizer (Supplementary Note 2; Fig. S7f-h). Finally, we tested different deep learning architectures including non-convolutional or recurrent long short-term memory (LSTM) networks. While very large networks tended to slightly overfit the data, most networks performed almost equally well (Supplementary Note 2; Fig. S8). Hence, the expressive power of moderately deep networks and the robustness of back-propagation with gradient descent enables multiple different networks to find good models for spike inference irrespective of the network architecture, hyper-parameter settings and the chosen learning procedure. This high robustness of the deep learning approach practically eliminates the need for manual adjustments of hyper-parameters by the user.

Generalization across neurons within the same dataset

Ideally, the ground truth data used to train a network should match the experimental conditions in the test dataset (calcium indicator type, labeling method, concentration levels, brain region, cell type, etc.). To explore spike inference under such conditions we measured how well spike rates of a given neuron within a ground truth dataset can be predicted by networks that were trained using the other neurons in the dataset. First, all ground truth calcium $\Delta F/F$ data were resampled to a common sampling rate and adjusted to the same noise levels by additive Poisson noise; if the initial noise level of a given ground truth neuron was higher than the target noise level, the neuron was not considered for the respective noise level analysis. We then evaluated the performance of CASCADE as a function of the noise levels of the (re-sampled) datasets. As expected, correlations increased and errors decreased for lower noise levels, while average biases seemed not to be systematically affected (Fig. 3a-d). Performance metrics also varied considerably for different neurons within a single dataset when resampled at the same noise level ν. To better understand this variability, we performed additional analyses.

First, we found spike-evoked calcium transients to be variable across neurons from the same dataset (Fig. S1). Large errors and biases, as well as low correlations, were observed when spike-evoked calcium transients of a neuron deviated strongly from those of other neurons (red arrow in Fig. 3d; cf. Fig. S1q for the respective linear kernels of DS#17).

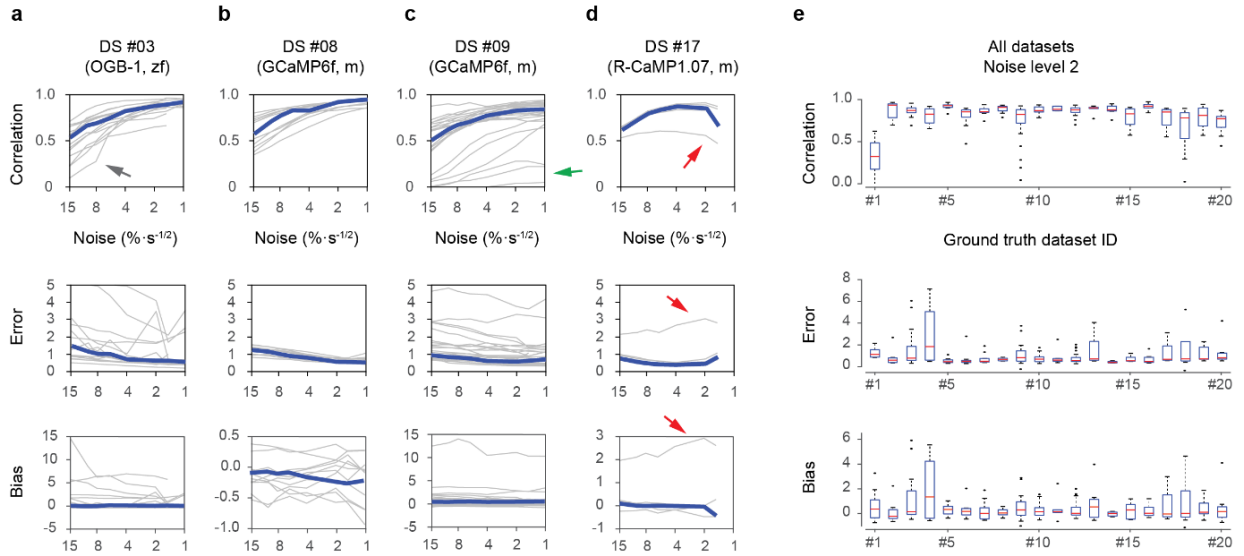

Figure 3 | Generalization across neurons within a dataset. The deep network was trained on all neurons of a specific dataset except one, and then tested with the remaining neuron. This analysis shows how the network is able to generalize to new neurons recorded under the same conditions, as a function of noise level ν. **a,** Performance of the predictions for 4 selected ground truth datasets in terms of correlation, error and bias as a function of the noise level. Error values were cropped at a value of 5 for display purposes. Single neurons in grey, median across neurons in blue. Grey lines highlighted by arrows indicate outlier neurons with low particularly low spike rates (black and green arrows) and particularly distinct calcium response kernel (red arrow, see main text for discussion). **b,** Correlation, error and biases as a distribution across neurons within each dataset, at a noise level of 2. All datasets were re-sampled at a frame rate of 7.5 Hz.

Second, spike inference may be complicated by movement artifacts or neuropil contamination. Movement artifacts typically had slow onset- and offset-kinetics (Fig. S9a), or a faster, quasi-periodic temporal structure related to breathing (Fig. S9d-e). Neuropil contamination is often very difficult to distinguish from somatic calcium signals and particularly severe when neurons are tightly packed and densely labeled[1,36–39] (Fig. S9b). For a subset of datasets, we tested the effect of simple center-surround subtraction of the neuropil signal[25]. Because subtraction is not perfect, decontaminated datasets still contained residual neuropil signals or negative transients (Fig. S9c). Nonetheless, spike inference was significantly improved by neuropil decontamination (Fig. S10). More detailed inspection of the results showed that CASCADE was able to learn to ignore negative transients and movement artifacts, but only as long as they were distinguishable from true calcium transients. For example, artifacts were only sometimes correctly not associated with spikes (Fig S9a,b,c), limiting the precision of spike inference.

Third, we found that the activity of sparsely firing neurons is less well predicted since the calcium signal of single action potentials is more likely to be overwhelmed by shot noise, particularly in the high-noise regime (arrows in Fig. 3a,c). We therefore evaluated conditions required for single-spike precision and observed that either shot noise or other noise sources were too prominent in all ground truth datasets to allow for reliable single-spike detection. The trained network thus systematically underestimated single spikes (Fig. S11). This observation was made using GCaMP

indicators, which show a strongly nonlinear relationship between calcium concentration and fluorescence and therefore are less sensitive to isolated single spikes occurring during low baseline activity, but also using synthetic dyes (Fig. S11). These observations indicate that the network learns a tradeoff between false-positive detections of noise events and false-negative detections of single spikes. Further details related to single-spike precision and the possibility to discretize inferred spike rates are discussed in Supplementary Note 3.

In summary, we showed that CASCADE is able to generalize to unseen neurons from the same ground truth training set. Not surprisingly, the precision of this generalization decreases for higher noise levels, in particular if spike rates are low. The precision is fundamentally limited by the variability of calcium kernels across neurons, and it is reduced when additional noise (motion artifacts, neuropil contamination) is prominent.

Generalization across datasets

We next explored how spike inference by a network trained on one ground truth dataset generalizes to other datasets. Using all available datasets, we quantified the median performance metrics across all possible combinations of datasets for training and testing (Fig. 4a,c,e; rows 1-20) and analyzed the performance of each trained model across test datasets (Fig. 4b,d,f). In most training/test combinations, correlations were high whereas errors and biases remained low. However, models trained with some datasets showed low performance across datasets (e.g., DS#01), and some datasets could not be predicted well by any model trained on another dataset (e.g., DS#01, DS#20). The entries of the matrix in Fig. 4 remained highly similar when parameters such as the resampling rate or the noise level were modified (Fig. S12). Interestingly, the performance of training/testing combinations showed no obvious clustering related to indicator type (e.g., genetically encoded vs. organic indicators) or species (zebrafish vs. mouse). Near-maximal correlation for a given dataset was often achieved by multiple models. In some datasets, the highest correlation was achieved when the model was trained on ground truth from another dataset, rather than from the same dataset. An attempt to explain the mutual predictability of datasets by more refined statistical dataset descriptors was not successful (Fig. S13). It is therefore not obvious how to select an optimal training dataset to predict spike rates for an unseen dataset.

To optimize dataset selection and network training for practical applications we explored three different strategies. First, we averaged across predictions of all models (Fig. 4a; row labeled ‘Average model’). Second, we used only a specific selection of datasets (mouse datasets using either GCaMP6f or GCaMP6s) for training. Third, we trained a ‘Universal model’ on all datasets except DS#01 (and excluding the dataset chosen for testing). We found that the universal model performed better than all other options (Fig. 4a-f). Compared to randomly selecting a single dataset for training, correlations were increased by 0.07±0.04, errors were reduced by 0.36±0.35, and biases were reduced by 0.28±0.26 (mean ± s.d., for all differences). In addition, the universal model performed better than any of the 20 single models when evaluated for all datasets ($p <$

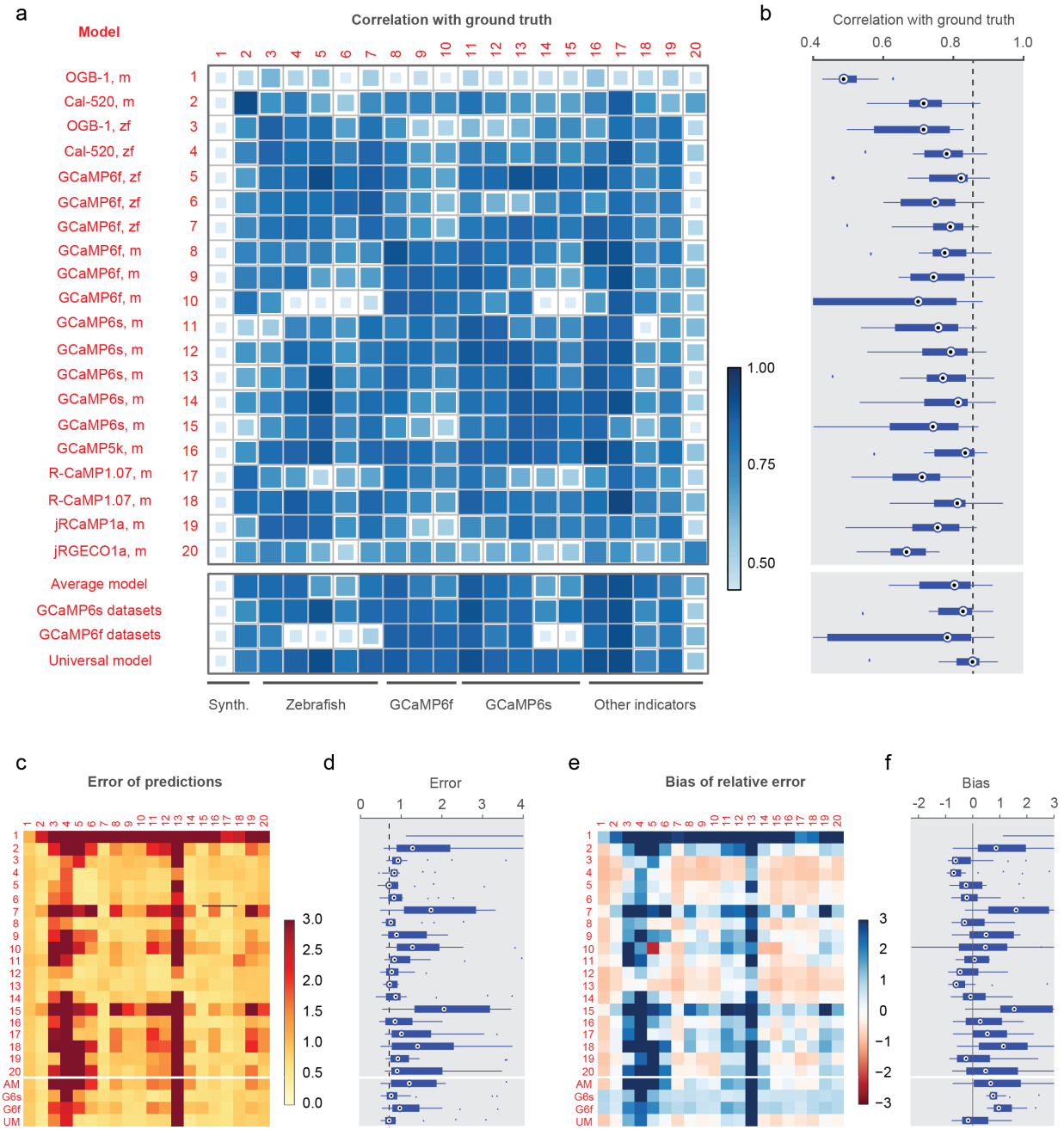

Figure 4 | Generalization across datasets. The network was trained on a given dataset (indicated by the row number) and tested on each other ground truth dataset (column). Diagonal values correspond to metrics shown in Fig. 3e. 'm': datasets recorded in mice, 'zf': zebrafish. Rows 21-24 are networks trained with a combination of several ground truth datasets (see text for more details). The 'universal model' is globally trained on all datasets except DS#01 and the test dataset. **a,** Correlation of predictions with the ground truth. The size of the squares scales with correlation. **b,** Distribution of the performance of each trained network (row) across all datasets. The dashed line highlights the median of the best-performing model ('universal model'). **c-d,** Relative error of predictions compared to the ground truth. The dashed line highlights the median of the best-performing mode ('universal model'). **e-f,** Relative bias of predictions compared to the ground truth. All datasets were re-sampled at a frame rate of 7.5 Hz, with a noise level of 2.

0.005 for all comparisons, paired signed-rank test). Compared to predictions across neurons within the same dataset (Fig. 3; diagonal elements in Fig. 4), the correlations resulting from the universal model were decreased by 0.01±0.02 (p = 0.24, Wilcoxon signed-rank test), errors were increased by 0.13±0.13 (p = 0.02), while the absolute bias was slightly decreased (0.11±0.31, p = 0.13). Hence, the performance of the universal model was similar to the performance of models with ground truth from the same datasets. Training the algorithm with all available reliable data is therefore a simple and effective strategy to generate a model that generalizes robustly to unseen datasets.

Comparison with existing methods

To benchmark the performance of CASCADE relative to alternative approaches we compared it to three other model-based methods: the Peeling algorithm for inference of discrete spikes[10], the fast online deconvolution procedure OASIS, which is contained in the CaImAn environment[14,38], and the more complex algorithm MLSpike, which in previous assessments outperformed various other methods[11,15]. Although model-based methods are, in principle, non-supervised, several parameters need to be tuned to achieve maximal performance on a given dataset[12]. We therefore used extensive grid searches to find the best parameters for each algorithm and each ground truth dataset (Methods; Table 2).

We determined the distribution of performance values (correlation, error, bias) across neurons for all combinations of methods and datasets (Fig. 5a-d; Figs. S14, S15). Overall, predictions of spike rates by CASCADE displayed higher correlations to the ground truth spike rates than predictions made by other algorithms (Fig. 5e; MLSpike: increase Δ=0.05±0.01, p < 1e-25; Peeling: Δ=0.12±0.01, p < 1e-40; OASIS: Δ=0.18±0.02, p < 1e-40; pseudomedian ± 95% CI, signed-rank test). Consistent with this result, predictions by CASCADE explained a greater fraction of the variance in the data than predictions by MLSpike, the best-performing model-based algorithm (Fig. 5f). The difference remained even when the MLSpike model was not optimized by the mean squared error but directly by the correlation (Δ=0.04±0.01, p < 1e-30). Interestingly, the unexplained variances were highly correlated, indicating that the errors made by CASCADE and MLSpike were qualitatively similar (Fig. 5g; see also red arrow heads in Fig. 5a and Fig. S15). This observation suggests that there are limitations to spike inference that are inherent to the data and cannot be overcome by any available algorithm, consistent with previous findings[15].

Next, we compared the generalization of CASCADE and MLSpike to unseen datasets. While we used the universal model for CASCADE (see Fig. 4a), MLSpike was tuned to obtain maximal performance across all datasets. In addition, for MLSpike, standard saturating and sigmoid nonlinearities were activated for synthetic dyes and genetically encoded sensors, respectively (see Methods). Furthermore, MLSpike was separately tuned for mouse datasets using GCaMP6f/s (DS#08-10 and DS#11-15, respectively) to account for prior knowledge about indicators. Hence, the model of MLSpike used here is more complex than the universal CASCADE model and, strictly speaking, does not perform true out-of-dataset generalization. Still, we obtained significantly higher correlation performance with CASCADE as compared to MLSpike

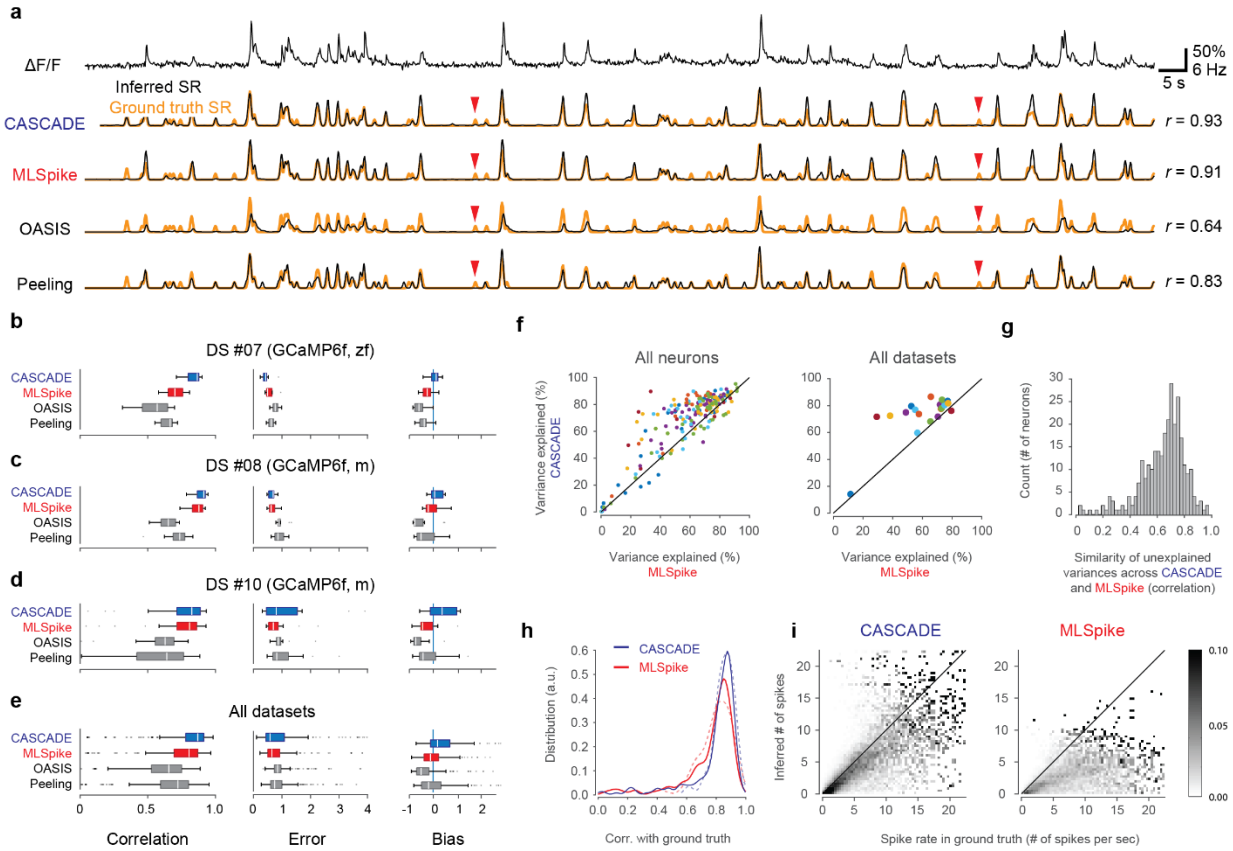

Figure 5 | Comparison with model-based algorithms. a, Example predictions from the deep-learning based method (CASCADE) and three model-based algorithms (MLSpike, OASIS, Peeling) of a ΔF/F recording. Respective predictions are in black, ground truth in orange. *r* indicates correlation of predictions with ground truth. Clear false negative detections are labeled with red arrowheads. **b-e,** Comparison of the four algorithms when optimized for a single dataset, with correlation (left), error (middle) and bias (right). **f,** Variance explained (r^2) by the MLSpike algorithm vs. variance explained by CASCADE. Each color corresponds to a ground truth dataset. **g,** Errors made, quantified as the unexplained variance, are highly corrected between MLSpike and CASCADE predictions. **h,** Distribution of correlation values across single neurons for CASCADE (black) and MLSpike (red) when applied to unseen datasets. Dashed lines show distributions for GCaMP in mice dataset only, for which MLSpike was specifically fine-tuned (see text for details). **i**, Spiking activity in 2 s-bins of ground truth vs. predictions for CASCADE (left) and MLSpike (right) as a heat map. The heat map has been normalized for both rows and columns to highlight the less frequently occurring spiking activity bins. All quantifications have been performed with ground truth datasets resampled at 7.5 Hz with a noise level of 2.

across all neurons (Fig. 5h; correlations 0.85±0.19 vs. 0.82±0.20, median ± s.d., for CASCADE vs. MLSpike; p = 0.0016, Wilcoxon signed-rank test), even when only those GCaMP6 datasets were considered, for which MLSpike was optimized (Fig. 5h, dashed lines). For errors and biases, no significant differences were found (p = 0.28 for errors and p = 0.12 for biases; distributions not shown). Together, our analysis indicates that CASCADE performs significantly better than the optimized MLSpike model not only for predictions within datasets, but also when applied to unseen data.

Interestingly, all model-based algorithms were found to be biased towards underestimating true spike rates, in particular OASIS (Fig. 5b-e, right column). To better understand this bias across activity regimes, we plotted the number of spikes for ground truth and predictions within each 2-s time bin (Fig. 5i; Fig S16 for OASIS and Peeling). The 2d histogram emphasizes the clear tendency of MLSpike, OASIS and the Peeling algorithm to underestimate larger events, which was much less pronounced for CASCADE.

Finally, we investigated the computation time required to perform predictions and found that CASCADE running on a GPU processed >200,000 samples per second, outperforming MLSpike (approximately 800 samples per second) by more than two orders of magnitude and only being surpassed by OASIS (approximately 350,000 samples), which is specifically designed for high processing speed. CASCADE running on a standard CPU (approximately 20,000 samples per second) also outperformed both the Peeling (approximately 5,500 samples) and the MLSpike (800 samples) algorithms.

Application to large-scale population calcium imaging datasets

A transformation of calcium signals into estimates of spike rates may be desired for multiple reasons. First, the reconstruction of spike rates can recover fast temporal structure in neuronal activity patterns that is obscured by slower calcium signals. Second, calcium signals contain shot noise and potentially other forms of noise that are unrelated to neuronal activity. A method that infers spike rates while ignoring noise can therefore de-noise activity measurements without compromising temporal resolution. Third, while calcium signals usually represent relative changes in activity, spike rates provide absolute activity measurements that can be compared more directly across experiments. With these potential goals in mind we applied CASCADE to different large-scale calcium imaging datasets.

In a brain explant preparation of adult zebrafish[22], we measured odor-evoked activity in the posterior part of telencephalic area Dp (pDp), the homolog of piriform cortex, using OGB-1. Multi-plane two-photon imaging[40] was performed with the same recording conditions as in DS#03 at a noise level of 2.36±0.97 (%·$s^{1/2}$; median ± s.d.) across 1,126 neurons. Under these conditions, predictions are expected to be highly accurate (Fig. 3a,e; correlation to ground truth: 0.87±0.06 for a noise level of 2, median ± s.d.). Consistent with previous electrophysiological recordings[41], spiking activity estimated by CASCADE was sparse (0.6±1.1 spikes during the initial 2.5 s of the odor response; mean ± s.d.; Fig. 6a) and variable across neurons (Fig. 6b), with a clear difference between the anatomically distinct dorsal and ventral regions of pDp (0.07±0.11 vs. 0.21±0.11 Hz; entire recording).

The comparison of ΔF/F signals and inferred spike rates showed that CASCADE detected phases of activity but effectively suppressed small irregular fluctuations in activity traces, indicating that spike inference suppressed noise. Consistent with this interpretation, spike inference by CASCADE increased the correlation between time-averaged population activity patterns evoked by the same odor stimuli in different trials (Fig. 6c,d). Previous studies showed that odor-evoked

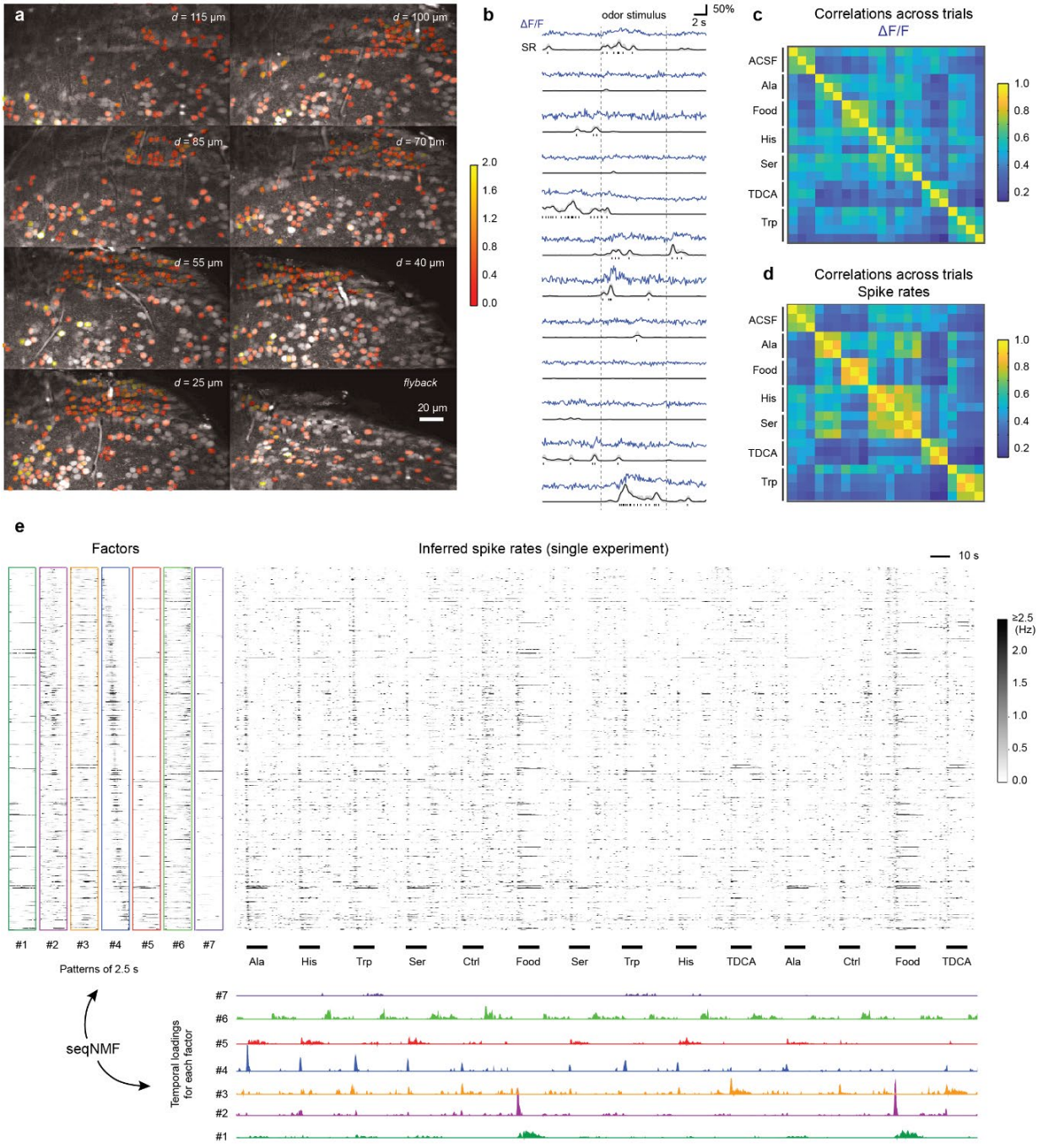

Figure 6 | Inference of spiking activity with CASCADE from population calcium imaging across >1100 neurons in adult zebrafish. a, Multiple planes were imaged simultaneously. The ROIs are colored with the average number of inferred stimulus-evoked spikes (colorbar). Non-active neurons were left uncolored. **b,** Randomly selected examples of calcium traces (ΔF/F, blue), inferred spiking probabilities (FR, black) and inferred discrete spikes, highlighting the de-noising through spike inference. **c,** Correlation of odor-evoked responses across trials, based on ΔF/F data during the initial 2.5 s of the odor response. **d,** Correlation of odor-evoked responses across trials, based on inferred spiking probabilities. **e,** Unsupervised detection of sequential factors (left) and their temporal 'loading' (bottom), shown together with the inferred spiking probabilities (center) across a subset of stimulus repetitions. The temporal loadings indicate when a given factor becomes active. All neurons were ordered according to highest activity in pattern #4, highlighting the sequential activity pattern that is evoked by stimuli at multiple times.

population activity in pDp is dynamic[41,42] but the fine temporal structure has not been explored in detail. We applied an unsupervised non-negative matrix factorization method for sequence detection (seqNMF, ref. [43]) to the inferred spike rate patterns to identify recurring short (2.5 s) sequences of population activity (factors) embedded in the overall population activity. This analysis required a high effective temporal resolution because factors exhibited rich temporal structure on a sub-second timescale (Fig. 6e). We found multiple factors that occurred with high precision and in a stimulus-specific manner at different phases of the odor response. For example, factors #2 and #4 in Fig. 6e were transient and associated with response onset, factor #5 persisted during odor presentation, and factor #6 was activated after stimulus offset (Fig. 6e). Odor-evoked population activity in pDp therefore shows complex dynamics on timescales that cannot be resolved without temporal deconvolution of calcium signals. The transformation of calcium signals into spike rate estimates by CASCADE therefore provides interesting opportunities to use calcium imaging for the analysis of fast network dynamics.

Next, we analyzed the Allen Brain Observatory Visual Coding dataset, comprising >400 experiments in mice with transgenic GCaMP6f expression, each consisting of approximately 100-200 neurons recorded at very low noise levels (0.94±0.25 %·$s^{-1/2}$; mean ± s.d.; Fig. 7a) [44]. Using CASCADE we estimated the absolute spike rates across all neurons for different transgenic lines (Fig. 7b,c; Fig. S17). Given the sampling rate (30 Hz) and noise level of this dataset we expect a correlation of 0.89±0.18, an error of 0.70±0.96 and a bias of 0.27±1.00 (median ± s.d. across neurons) based on our previous cross-dataset comparisons (Fig. 4). Since specific ground truth for interneurons is lacking, we did not include interneuron experiments in our analysis. The spike rates across the full recordings of all 40,055 neurons were well described by a lognormal distribution centered around 0.1-0.2 Hz (Fig. 7d). Across areas, spike rates were lowest in area RL compared to the other visual areas (Fig. 7e). Activity also varied systematically across cortical layers, with highest activity in layer 5 (Fig. 7c). These differences are unlikely to reflect cell type-specific biases in spike inference because region- and layer-specific differences were observed also within the same transgenic lines. Across all layers, activity during presentation of stationary or moving naturalistic stimuli (natural scenes or movies) appeared to be higher than responses to stationary or moving gratings (Fig. 7f). These results provide a comprehensive description of neuronal activity in the visual system of the mouse and reveal systematic differences in neuronal activity across cell types, brain areas, cortical layers and stimuli.

Raw ΔF/F often exhibited correlated noise, visible as a vertical striping in matrix plots, which was small for individual neurons but tended to dominate the mean ΔF/F across neurons, possibly due to technical noise or neuropil signal (Fig. 7g). CASCADE effectively eliminated these artifacts (Fig. 7h). As a consequence, correlations between activity traces of different neurons were reduced across all experiments by 38±43% (mean ± s.d.; Fig. 7i; $p < 1e\text{-}15$, paired signed-rank test). Many analyses of neuronal population activity require accurate measurements of pairwise neuronal correlations[44–47]. Noise suppression by spike inference can therefore help to make these analyses more reliable.

Together, these examples illustrate how calibrated spike inference by CASCADE can be applied to perform comprehensive analyses of neuronal activity, to identify complex temporal structure in neuronal population dynamics, and to remove shot noise and other noise from calcium imaging data.

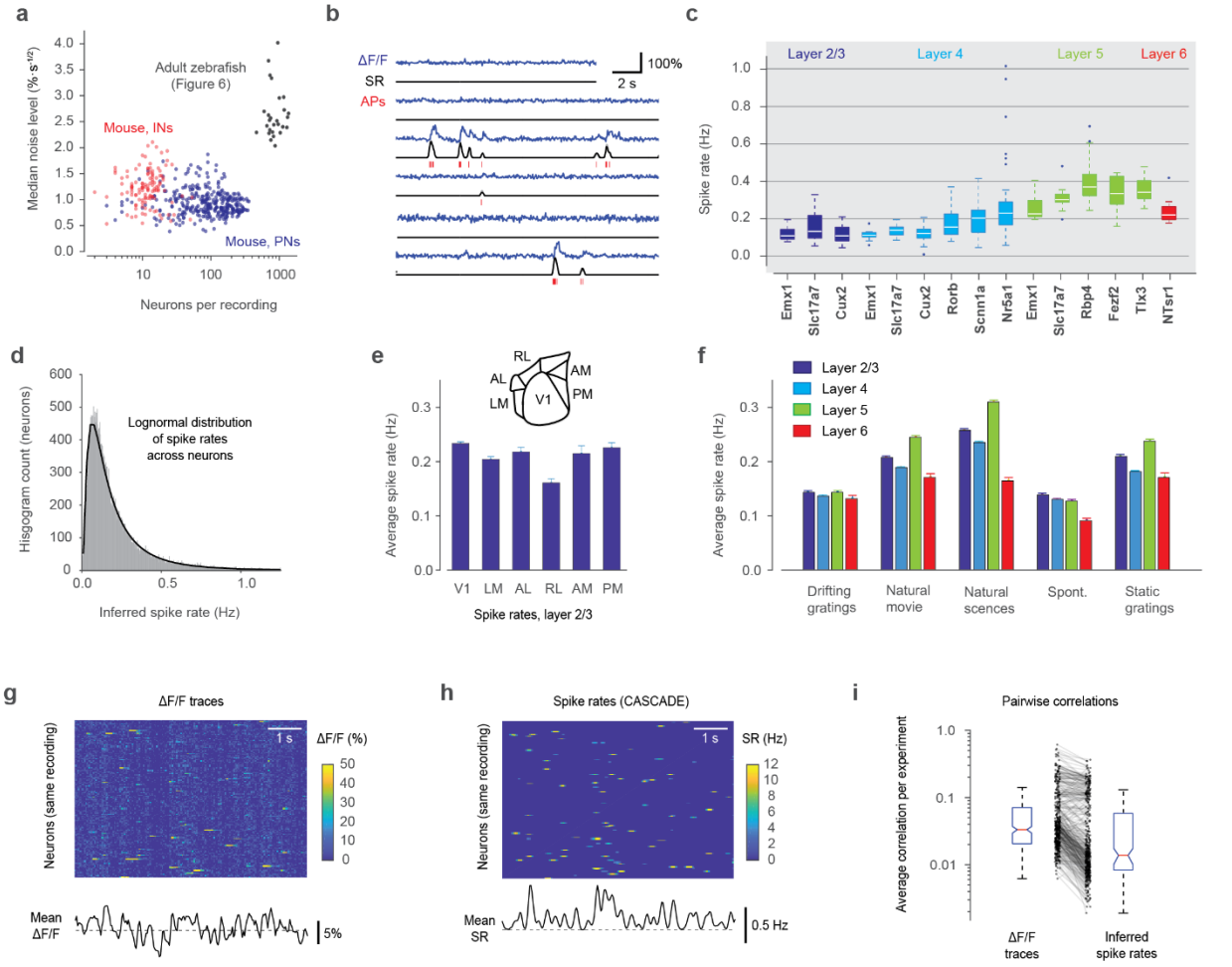

Figure 7 | Inference of spiking activity with CASCADE for the Allen Brain Observatory dataset in mice. a, Number of recorded neurons vs. median noise levels for all experiments from dataset from excitatory (blue) and inhibitory (red) datasets; population imaging datasets in zebrafish (Fig. 8) in black for comparison. **b,** Example predictions from calcium data (blue). Discrete inferred spikes are shown in red below the inferred spike rates (black). See Fig. S17 for more examples. **c,** Inferred spike rates across all neurons for recordings in different layers (colors) and for different transgenic driver lines of excitatory neurons. Each underlying data point is the mean spike rate across an experiment. Total of 336 experiments. **d,** Spike rates across the entire population are well described by a log-normal distribution (black fit). n = 38‘466 neurons**. e,** Spike rates across the visual areas, shown for all L2/3 experiments (primary area V1, lateral area LM, anterolateral area AL, posteromedial area PM rostrolateral area RL, anteromedial area AM). **f,** Average spike rates for different stimulus conditions (x-labels) across layers (colors). **g** Excerpt of raw ΔF/F traces of a subset of neurons of a single experiment (L2/3-Slc17a7, experiment ID ‘652989705’). Correlated noise is visible as vertical striping patterns. **h,** Same as (g), but with inferred spike rates. **i,** Average correlation between neuron pairs within an experiment, computed from raw ΔF/F traces (left) and inferred spike rates (right).

A user-friendly toolbox for spike inference

The deployment of spike inference tools often suffers from practical problems. First, the difficulty to set up a computational pipeline might prevent wide-spread use. To address this problem, we generated a cloud-based solution using Colaboratory Notebooks that can be applied without local installations. For more engaged users, we set up a well documented Github repository (https://github.com/HelmchenLabSoftware/Cascade) containing ground truth datasets, pretrained models, notebooks and demo scripts that can be easily integrated into existing analysis pipelines like CaImAn, SIMA or Suite2P[38,48,49]. Since the algorithm works on regular laptops and workstations without GPU support, the main installation difficulties of typical deep learning applications are circumvented.

In a typical workflow, first the noise level for each neuron in a calcium imaging dataset is determined. Then, a pre-trained model that has been trained on noise-matched resampled ground truth data is loaded from an online model library and applied to the ΔF/F data without need to adjust parameters. In addition, CASCADE can be easily modified and retrained to address additional specific needs, for example, more complex loss functions[21] or a modified architectures. Also, the resampled ground truth can be adapted directly if desired. For example, we used a Gaussian kernel to smooth the ground truth spike rate, but this standard procedure can be disadvantageous to precisely determine the onset timing of discrete events. In CASCADE, it is simple to replace the Gaussian kernel by a causal smoothing kernel to circumvent this problem (Fig. S18).

A second problem is that experimenters may need additional tools and documentation for interpretation of the results. To address this issue, we included graphical outputs and guiding comments that are accessible also for non-specialists throughout the demo scripts. Together with existing literature on the appropriate interpretation of raw calcium data[4,5,21,50,51], this will help to focus the attention on data quality and the potentials and limitations of raw and deconvolved data.

DISCUSSION

We have created resources and tools for spike inference from unseen calcium imaging data. Any spike inference approach, in particular methods based on deep learning, critically depend on the availability and quality of ground truth datasets. We therefore created a ground truth database that is larger and more diverse than previous datasets[15,18], spanning various calcium indicators, brain regions and species (Fig. 1). Moreover, we developed CASCADE, a novel algorithm for spike inference based on deep learning. The central idea of CASCADE is to optimize the match between the training data and experimental datasets, rather than to optimize the inference algorithm itself. Previous supervised spike inference algorithms typically trained their model with an existing, immutable ground truth[15,18,19]. Our algorithm, in contrast, resamples the ground truth datasets upon demand to match both frame rate and noise level in an automated fashion for each neuron (Fig. 2). Training with an appropriately resampled and noise-matched ground truth

resulted in substantially improved inference, highlighting the importance of training the algorithm not only with realistic calcium signals but also with realistic noise patterns.

The generalization of spike inference methods across unseen datasets had been investigated sporadically[12,13,20] but never systematically in previous studies, presumably due to the lack of extensive ground truth data. Taking advantage of our diverse ground truth database, we explored how predictions depend on species (zebrafish or mouse), indicator type and brain region (Fig. 3,4) and other experimental parameters that are presumed to strongly influence spike inference. Surprisingly, we found that some training datasets allowed for efficient generalization across these parameters, and a combined training dataset achieved very high performance across all test data. Moreover, some datasets performed poorly as training sets while others performed poorly as test sets, even when compared against datasets with a similar indicator and/or from the same brain region. These observations suggest that generalization is affected significantly by experimental differences that are difficult to identify, such as indicator concentration or baseline calcium concentrations. However, this problem could be overcome by training networks on a diverse ground truth database, indicating that networks can learn to take these variations into account when sufficient information is provided during training. Highly efficient generalization was obtained by the unsupervised 'universal' model that was trained on all available ground truth datasets. This universal model is therefore well-suited for practical applications of spike inference in unseen datasets.

In all investigated situations, our algorithm outperformed existing approaches (Fig. 5). Predictions were not only more precise, as measured by correlation metrics, but also less biased towards underestimates of true spike rates. We reason that the balance between spike detection and noise suppression is crucial for reliable spike inference. Our results suggest that less expressive models have to over-suppress if specific suppression is not possible. In contrast, the expressiveness of the employed deep networks enables CASCADE to better distinguish signal from noise, while their relatively small size prevents overfitting.

CASCADE was not sensitive to user-adjustable hyper-parameters or the class of the deep networks tested. This insensitivity has two consequences. First, future work should not necessarily focus on further improvements in model architecture but rather on the acquisition and processing of ground truth. Second, because hyper-parameters do not need to be adjusted by the user, the application of spike inference becomes simple in practice. While some previous studies assumed that user-adjustable parameters in model-based algorithms increase the interpretability of the model[10–13], we argue here that (1) biophysical model parameters are often ambiguous[12] and therefore not directly interpretable, and (2) it is more important to focus on the interpretability of the results rather than the model. To this end, our toolbox provides methods to estimate the errors made during spike inference. Moreover, we included a detailed documentation in the Colaboratory Notebook to help the user interpret the results.

Quantitative inference of spike rates is critical for the analysis of existing and future calcium imaging datasets[4,5,21]. Although ΔF/F can in theory only report on spike rate changes, we found that absolute spike rates can be reliably inferred when the baseline activity is sufficiently sparse,

which was the case in all datasets examined here (Fig. 6,7). The enhanced temporal resolution will be particularly useful for the analysis of neuronal activity during natural stimulus sequences and behaviors which occur on timescales shorter than typical durations of calcium transients. For example, the deconvolution of calcium signals can increase the effective temporal resolution to timescales below 100 ms, which will allow for the analysis of neuronal representations across theta cycles[52] and for the resolution of early and late dynamics in cortical responses to sensory inputs that have been associated with different processing steps[53]. Moreover, the inference of absolute spike rates will help improve the calibration of precisely patterned optogenetic manipulations[54–56] and the extraction of constraints, e.g. absolute spike rates, for computational models of neural circuits.

The reliability of spike inference obviously depends on the recording quality of the calcium imaging data. To improve data quality of ΔF/F signals, future work should focus on the reduction of movement artifacts and neuropil contamination by both experimental design[50,57] and extraction methods[36–39], including the correct estimation of the F_0 baseline despite unknown background fluorescence. In the long term, the development of more linear calcium indicators[58] and especially the acquisition and integration of more specific ground truth, *e.g.*, for inhibitory interneuron subtypes[59], will enable quantitative spike inference for an even broader set of experimental conditions. We envision that our set of ground truth recordings will become enlarged over time, allowing to train more and more specific models for reliable inference of spike rates.

ACKNOWLEDGEMENTS

We thank the members of the GENIE project, the Allen Institute and the Spikefinder project for publicly providing existing ground truth datasets together with excellent documentation. We thank Philipp Berens and Emmanouil Froudarakis for providing additional information on the Spikefinder datasets. We thank Gwendolin Schoenfeld for helpful discussions on dataset 17, and Hendrik Heiser, Nesibe Temiz, Chie Satou, Gwendolin Schoenfeld and Henry Luetcke for testing earlier versions of the toolbox. This work was supported by grants to F.H. from the Swiss National Science Foundation (Project grant 310030-127091; Sinergia grant CRSII5-18O316) and by the European Research Council (ERC Advanced Grant BRAINCOMPATH, grant agreement no. 670757), by grants to K.K. from MEXT, Japan (Scientific Research for Innovative Areas, no. 17H06313), by grants to R.W.F. from the Swiss National Science Foundation (Project grant 310030B-152833/1) and from the European Research Council (ERC Advanced Grant MCircuits, grant agreement no. 742576), by the Novartis Research Foundation, and by a fellowship from the Boehringer Ingelheim Fonds to P.R..

CONTRIBUTIONS

P.R. conceived the project, developed the algorithm, performed ground truth recordings (datasets 3-7), performed all analyses, developed the toolbox and wrote the paper. S.C. performed ground

truth recordings (datasets 17 and 18). A.H. developed the toolbox. M.E. and K.K. performed ground truth recordings (dataset 2). F.H. supervised ground truth recordings (datasets 17 and 18) and the development of the toolbox, and wrote the paper. R.W.F. supervised ground truth recordings (datasets 3-7) and the development of the algorithm, and wrote the paper.

COMPETING INTERESTS

The authors declare no competing interests.

METHODS

Simultaneous juxtacellular recordings and calcium imaging in adult zebrafish

All zebrafish experiments were approved by the Veterinary Department of the Canton Basel-Stadt (Switzerland). For the recordings in DS#03 and DS#04, the adult zebrafish brain was dissected *ex vivo* as described before[22] and OBG-1 AM or Cal-520 AM were injected and incubated in posterior Dp (pDp) as described before[62]. During the dissection, the dura mater above pDp was carefully removed to prevent clogging of the patch pipette. Two injections each took 1-2 min (injection 1: ~210 µm dorsal from the ventralmost aspect of Dp and ~130 µm from the lateral surface of Dp; injection 2: 180 µm and 60 µm). Dye injection was monitored by snapshot multiphoton images, and pressure was adjusted to avoid fast swelling of the tissue.

Juxtacellular recordings were performed >1h and <4h after the dye injection. Patch pipettes were pulled from 1 mm thin borosilicate glass capillaries (Hilgenrath), with a pipette resistance of 5-8 MΩ. Micropipettes were backfilled with ACSF (in mM: 124 NaCl, 2 KCl, 1.25 KH_2PO_4, 1.6 $MgSO_4$, 22 D-(+)-Glucose, 2 $CaCl_2$, 24 $NaHCO_3$; pH 7.2; 300-310 mOsm) with 0.05 mM Alexa 594.

The explant preparation was rotated about the anterior-posterior axis to allow for optical access from the side (sagittal imaging). Two-photon imaging together with transmitted PSD imaging were used to target the pipette to pDp, while continuous low pressure (30-40 mBar) was applied to prevent clogging of the pipette. The pipette then entered the tissue with initial high pressure (90-110 mBar) that was lowered after a few seconds. Neurons were approached using the shadow-patching technique described previously[42,63], but with lower pressure applied. Juxtacellular recordings were performed after establishing a loose seal with a target neuron (typically 30-50 MΩ seal resistance). In some cases, a light constant negative pressure was applied initially to improve the electrical contact with the target cell. In several cases, single micropipettes were used multiple times. Recordings were performed in voltage-clamp mode with the voltage adjusted such that the resulting current approximated zero, following the procedures described elsewhere[64].

For DS#05-07, which were based on a transgenic line expressing GCaMP6f in the forebrain[40], the experimental procedures were similar but did not require injection and incubation of synthetic dyes. The main difference from recordings with OGB-1 or Cal-520 was the lower baseline brightness for GCaMP6f, which made it often more difficult to identify neurons. Upon application of odor stimuli, stimulus-responsive neurons that expressed GCaMP6f became brighter, which permitted reliable visual identification of neurons for targeted patching. For neurons in dD (DS#06) with no obvious odor responses, random cells were patched based on shadow images generated by the blown-out Alexa dye[63].

Simultaneous recording of fluorescence and extracellular spiking of the same neuron was synchronized using *Scanimage 3.8* for imaging[65] and *Ephus* for electrophysiology[66]. Calcium imaging was performed at intermediate zoom with a framerate of 7.5 or 7.8125 Hz for DS#03 and DS#04 and at high zoom with a framerate of 30 Hz for DS#05-07. Electrophysiological recordings were lowpass-filtered at 4 kHz (4-pole Bessel filter) and sampled at 10 kHz.

Recordings were performed in 120-s episodes, with repeated stimulation of the nose with food extract odorants as described previously[42]. For recordings in pDp, spike rates are often very low (or zero). When no spiking activity was observed, the holding potential of the pipette was set to higher values (between +5 and +30 mV), leading to an extracellular current that depolarized the neuron if the seal resistance was sufficiently high, resulting in artificially generated spikes. If no spikes could be elicited over the full duration of the recording, the recording was not included in the ground truth dataset.

Anatomical location and further information about neurons in zebrafish ground truth datasets

DS#03: OGB-1, injected in the posterior part of the olfactory cortex homolog (pDp) in adult zebrafish. Recordings were performed throughout dorsal and ventral compartments of pDp and OGB-1 was injected as described previously[62]. Because OGB-1 localizes predominantly to the nucleus and because the resolution was high, neuropil contamination is negligible in this dataset.

DS#04: Cal-520, injected in the posterior part of the olfactory cortex homolog (pDp) in adult zebrafish. Same brain region as DS#03. Unlike OGB-1, Cal-520 is primarily cytoplasmic, resulting in considerable neuropil contamination. Cal-520 spread less than OGB-1 after injection and labeled only a small central volume in pDp.

DS#05: tg(NeuroD:GCaMP6f), anterior part of the olfactory cortex homolog (aDp) in adult zebrafish. In this transgenic line, GCaMP6f is strongly expressed throughout Dp. Recording location and framerate were chosen to match previous experiments[42].

DS#06: tg(NeuroD:GCaMP6f), dorsal part of the dorsal pallium (dD) in adult zebrafish. All recorded neurons in dD were mapped onto brain regions Dm, Dl, rDc and cDc based on neuroD expression in the dorsal part of the dorsal pallium (Fig. S19, following Huang et al., 2020, ref. [67]). Although the dorsal pallium is not known to be directly involved in olfactory processing, we noticed that several neurons were clearly inhibited during odor stimulations (duration, 10 - 30 s).

DS#07: tg(NeuroD), olfactory bulb (OB) in adult zebrafish. In the olfactory bulb of this transgenic line, GCaMP6f is restricted to a distinct, small subset of putative mitral cells and interneurons[40]. Neurons #1-#3, #5 and #7 were identified as interneurons based on their small size and morphology, while neurons #4, #6, #8 and #9 were classified as putative mitral cells.

Simultaneous juxtacellular recordings and calcium imaging in anesthetized mice (R-CaMP1.07)

All experimental procedures were conducted in accordance with the ethical principles and guidelines for animal experiments of the Veterinary Office of Switzerland and were approved by the Cantonal Veterinary Office in Zurich. For virus-induced expression of R-CaMP1.07, AAV1-EFα1-R-CaMP1.07 and AAV1-EFα1-DIO-R-CaMP1.07 were stereotactically injected under isoflurane anaesthesia in barrel cortex of C57BL/6J mice and hippocampal area CA3 of tg(Grik4-

cre)G32-4Stl mice, respectively (300 nL with approximately $1x10^7$ vg/nL)[23]. We performed combined electrophysiology and in vivo calcium imaging in acute experiments in anesthetized animals (n = 3; at least two weeks after virus injection) as described previously for barrel cortex recordings[23]. A stainless steel plate was fixed to the exposed skull using dental acrylic cement. A 1x1 mm^2 craniotomy was performed over barrel cortex. The dura was cleaned with Ringer's solution (containing in mM: 135 NaCl, 5.4 KCl, 1.8 $CaCl_2$, 5 HEPES, pH 7.2 with NaOH) and carefully removed. To reduce tissue motion caused by heart beat and breathing, the craniotomy was filled with low concentration agarose gel and gently pressed with a glass coverslip. For CA3 recordings, a 4 mm Ø craniotomy was centred over the virus injection locus. The overlying cortex was aspirated until the corpus callous became visible. 1% agarose gel was filled into the cavity to reduce tissue motion.

Juxtacellular recordings from R-CaMP1.07-expressing neurons were obtained with glass pipettes (4–6 MΩ tip resistance) containing Ringer's solution. For pipette visualization Alexa-488 (Invitrogen) was added to the solution or pipettes were coated with BSA Alexa-594 (Invitrogen). Action potentials were recorded in current clamp at 10 kHz sampling rate using an Axoclamp 2B amplifier (Axon Instruments, Molecular Devices) and digitized using Clampex 10.2 software.

Simultaneous juxtacellular recordings and calcium imaging in anesthetized mice (DS#02, Cal-520)

C57BL6/J male mice (postnatal days 28–61) were anesthetized by intraperitoneal injection of 1.9 mg/g urethane and the skull was partly exposed and attached to a stainless steel frame, as described before[24]. In a small craniotomy of the barrel cortex (diameter ~2 mm; ±3 mm lateral and -1.5 mm caudal from bregma), we removed the dura, filled the cranial window with 1.5% agarose and placed a coverslip over the agarose to minimize brain movements, as described before[24]. Cal-520 AM together with an Alexa dye were bolus-loaded in layer 2/3 of the barrel cortex (200–300 µm deep below the surface) and monitored by two-photon imaging on the Alexa channel, as described before[24]. Calcium imaging was performed more than 30 min after dye ejection.

For simultaneous calcium imaging and loose-seal cell-attached recordings, we filled glass pipettes (5–7 MΩ) with the extracellular solution containing Alexa 594 (50 µM), inserted pipettes into the barrel and targeted somata that had been loaded with Cal-520 AM. At about 10 min after the establishment of the cell-attached configuration, we performed simultaneous loose-seal cell-attached recording and high-speed line-scan calcium imaging (500 Hz) on the soma of cortical neurons. The electrophysiological data were filtered at 10 kHz and digitized at 20 kHz by using Multiclamp 700B and Digidata 1322A (Molecular Devices), and acquired by AxoGraph X (AxoGraph).

Analysis of simultaneous juxtacellular recordings and calcium imaging

Movies of calcium indicator fluorescence images were corrected off-line for movement artifacts, i.e., slow drifts due to relaxation of the brain tissue for zebrafish data or fast movement artifacts

for recordings in anesthetized mice. Ground truth recordings from DS#02 were not corrected for movement artifacts due to the scanning modality (line-scan). Afterwards, regions of interest (ROIs) were manually drawn using a custom-written software tool (https://git.io/vAeKZ)[42] for each trial to select pixels that reflected the calcium activity of the neuron. Fluorescence traces were extracted either as average across the ROI or individually for each pixel to allow for both natural and artificial sub-sampling of calcium signal noise levels (Fig. S4).

Spike times were extracted from juxtacellular recordings using a custom-written template-matching algorithm. In brief, peaks of the first derivative of a 1 kHz-filtered electrophysiological signal were detected using a threshold that differed between recordings and that was manually adjusted to safely exclude false positives. The original waveforms of the detected events were then averaged and used in a second step as a template to detect all events across the full recording more precisely via cross-correlation of the template with the original signal. A threshold adjusted manually for each recorded neuron extracted action potential events. The process of first generating a template that was afterwards used to detect stereotypic signals allowed to increase the signal-to-noise of detected events, similar to previous usages of template matching in electrophysiology[68,69].

Quality control

All electrical spiking events were inspected visually and compared to simultaneously recorded calcium transients. Any recordings that were ambiguous due to low electrophysiological signal-to-noise of action potentials were discarded and not used for any further analysis. In addition, neurons were discarded from analysis which did not spike even after application of currents, as well as neurons which became visibly brighter after establishing a loose seal due to unclear, possibly mechanical reasons. If the calcium recording clearly exhibited events without corresponding electrophysiological action potential, the calcium trace of the manually drawn ROI and the calcium traces of adjacent neurons or neuropil were inspected together with the electrophysiological recordings in order to assess optical bleed-through, and ROIs were corrected to avoid contamination. We also noted that occasionally mechanical stress exerted by the recording pipette can increase the brightness of the recorded neuron[26], possibly by the release of internal calcium stores. Recordings made during and after such events were discarded. Bursting lead to adaptation of the extracellularly measured spike amplitude. Such recordings (e.g., in DS#17 with bursts of >10 APs with an inter-spike interval of ca. 5 ms) were carefully inspected for missed low-amplitude action potentials, in particular during these bursts.

Extraction of ground truth from publicly available ground truth datasets

Additional ground truth was extracted from publicly available datasets and quality-controlled for each neuron[15,18,25,27–29].

The Allen Institute datasets

For DS#09-12 from Huang et al. (2020)[25], raw fluorescence traces were extracted from the processed datasets which were downloaded from https://portal.brain-map.org/explore/circuits/oephys. Neuropil was subtracted using the same standard scaling value for all neurons to make recordings comparable with other datasets (neuropil contamination ratio 0.7), despite the caveats associated with this procedure[25]. A 6-s running 10% lowest percentile window was typically used to compute F_0 for $\Delta F/F_0$ calculation, but percentile values were adjusted to the noisiness of the recording and over window durations that were adjusted to the baseline activity. Simultaneous juxtacellular and calcium imaging recordings were inspected for each ground truth neuron together with the raw movie. Recordings with excessive movement artifacts or apparent inconsistencies of juxtacellular and calcium recordings were discarded entirely.

The Spikefinder datasets

For DS#01, DS#14 and DS#15 from Theis et al. (2016)[18], the ground truth recordings at their native sampling rates as released during the Spikefinder challenge[15] were processed. This Spikefinder dataset consists of 5 separate datasets. Datasets 1 and 4 of these 5 datasets were excluded since fluorescence baseline and the scaling of fluorescence were unknown. The other datasets were extracted as fluorescence traces, F_0 was computed as the 10^{th} percentile value (adjusted depending on the firing rate of each neuron) and used to compute $\Delta F/F_0$. Some ground truth neurons were discarded due to a highly unstable calcium recording baseline, but no strict quality control was possible since the raw calcium imaging data were not available. As found during a previous study, some datasets of the Spikefinder challenge come with calcium recordings that are delayed with respect to the electrophysiological recordings[15]. We therefore manually corrected for delays of the calcium recording with respect to the electrophysiological recording based on visual alignment of extracted linear kernels. The same correction delay was applied across all neurons of a given dataset.

The GENIE datasets

Datasets DS#08, DS#013, DS#016, DS#19 and DS#20 were downloaded from http://crcns.org/data-sets/methods[27–29,70,71]. For DS#08 and DS#13[29], ROIs were extracted from raw calcium imaging data using the same approach as described above for R-CaMP1.07 data. Recordings with excessive movement artifacts or apparent inconsistencies of juxtacellular and calcium recordings were discarded entirely. Neuropil was subtracted using the same standard scaling value for all neurons (neuropil contamination ratio 0.7)[29]. F_0 values were computed using percentile values that were adjusted to the noisiness of the recording, and over window durations that were adjusted to the baseline activity.

For datasets DS#16, DS#19 and DS#20, no raw calcium imaging data were available, therefore not allowing for strict quality control using raw calcium recordings as additional feedback. Neuropil was subtracted from raw fluorescence using the same standard scaling value for all neurons (neuropil contamination ratio 0.7) [27,28]. F_0 values were computed using percentile values that were adjusted to the noisiness of the recording, and over window durations that were adjusted to the baseline activity.

Population calcium imaging with OGB-1 in zebrafish pDp

Ex vivo surgeries, OGB-1 AM injections and calcium imaging were performed as described above for juxtacellular recordings. Calcium imaging in Dp was performed using a custom-built multiphoton multiplane microscope based on a voice-coil motor for fast z-scanning as described before[40]. Laser power below the objective was 29-35 mW (central wavelength 930 nm, temporal pulse width below the objective 180 fs), with higher laser power for deeper imaging planes.

Imaging was performed in 8 planes (256x512 pixels, ca. 100x200 µm) at 7.5 Hz over a z-range of approximately 100 µm. Due to slowly relaxing brain tissue, movement correction was applied every 5 min by acquiring local z-stacks with a z-range of ±6 µm. The maximum cross-correlation between a reference stack acquired before the experiment and the local z-stack indicated the optimal positioning which was targeted using the stage motors of the microscope.

For odor stimulation, amino acids (His, Ser, Ala, Trp; Sigma) were diluted to a final concentration of 10^{-4} M and bile acid (TDCA; Sigma) was diluted to 10^{-5} M in ACSF, immediately before the experiment; and food extract was prepared as previously described[42]. Odors were applied for 10 s through a constant stream of ACSF using a computer-controlled peristaltic pump[42] in a pseudo-random order with three repetitions of each odor presentation.

Extraction of linear kernels from ground truth data

To extract linear kernels, we used simple regularized deconvolution using the *deconvreg*(*Calcium*,*Spikes*) function in Matlab (Mathworks). This function computes the kernel which, when convolved with the observed *Spikes*, results in the best approximation of the *Calcium* trace. Linear kernels were similar on average when extracted using different deconvolution algorithms (Wiener deconvolution, Lucy-Richardson algorithm; data not shown).

To compute the variability of linear kernels across neurons within and across datasets (Fig. S1), we split the ground truth recording of each neuron in five separate parts and computed the linear kernels for each of the segments separately. If the coefficient of variation (standard deviation divided by mean) across these five values was lower than 0.5, the computation of the kernel amplitude was considered reliable and included in the plots in Fig. S1.

Computation of noise levels

Shot noise is often the main factor determining the signal-to-noise level of calcium imaging approaches and can be quantified by the median absolute difference between two subsequent time points of the calcium ΔF/F trace. The median approximates the standard deviation but is more robust against large outliers including calcium transients generated by action potentials. We divided the median value by the square root of the frame rate f_r to obtain a measure for the noise level that is independent of the sampling rate. The square root was applied since this describes how shot noise decreases when the number of events increases ($SNR \sim \sqrt{N}$):

$$\nu = \frac{\mathrm{Median}_t|F_{t+1} - F_t|}{\sqrt{f_r}}$$

ν, when computed for ΔF/F data, is quantitatively comparable across datasets. A value of $\nu = 1$ will always be a very low level, while $\nu = 8$ will always be high. For the purpose of readability, we omitted units of ν throughout: $[\nu] = \% \cdot s^{-1/2}$.

Metrics to quantify performance of spike inference algorithms: Correlation, Error and Bias

The ground truth spike rates were generated from detected discrete spikes by convolution with a Gaussian smoothing kernel (except in Fig. S22, where a non-Gaussian, causal kernel was applied). The precision of the ground truth was adjusted by tuning the standard deviation of the smoothing Gaussian (typically σ = 0.2 s for 7.5 Hz recordings and σ = 0.05 s for 30 Hz recordings). The ground truth spike rate was then compared to the inferred spike rate.

There is no single metric to reliably reflect the goodness of performance of a spike inference algorithm. Correlation between the inferred spike rate and the ground truth is widely used[15], but fails to account for different absolute scaling or offsets. F_1-scores combine false positives and negatives[11] but are difficult to compare across datasets when the baseline spike rates vary (which is the case for our database). Other metrics try to combine the strengths of the correlation measure with a sensitivity to the correct number of spikes[72] but are less intuitive.

We defined three intuitive and complementary metrics. **Correlation** is a standard measure of the similarity between ground truth and inferred spike rates. The relative error (abbreviated as **error**) results from the sum of false positives and false negatives when subtracting the ground truth from the inferred spike rate, normalized by the absolute number of spikes in the ground truth. For example, an error of 0.7 would indicate that the number of either falsely inferred or falsely omitted spikes is about 70% of the number of spikes in the ground truth. Finally, the relative bias (or simply **bias**) is defined as the difference of false positives and false negatives, again normalized by the absolute number of spikes in the ground truth. Algorithms that systematically underestimate spike rates will tend towards the minimum of the bias, -1, whereas other algorithms may tend to systematically overestimate spike occurrences (bias > 0). Importantly, the error can be very high due to a high number of false positives and false negatives, but the bias might be still zero. Error and bias are therefore two metrics that describe the absolute errors in terms of spike rates, thus complementing the correlation metric. Figure S5 visually illustrates the definition of the error and bias metrics.

Architecture of the default convolutional network

The default network consists of a standard convolutional network with in total 6 hidden layers, including 3 convolutional layers. The input consists of a window of 64 time points symmetric around the time point for which the inference is made. The three convolutional layers have relatively large but decreasing filter sizes (31, 19, 5 time points), with an increasing number of

features (20, 30, 40 filters per layer). After the second and the third layer, maximum pooling layers are inserted. A final densely connected hidden layer consisting of 10 neurons relays the result to a single output neuron. While all neurons in hidden layers are based on rectified linear units (ReLUs), the output neuron is based on a linear identity transfer function. In total, the model consists of 18'541 trainable parameters.

The properties of the calcium imaging data are accounted for by resampling the ground truth with the appropriate noise levels and the matching frame rate. The ground truth is smoothed with a time-symmetric Gaussian kernel of standard deviation 0.2 s for resampling at 7.5 Hz and 0.05 s for 30 Hz or a causal kernel (inverse Gaussian distribution) to facilitate gradient descent.

Training deep networks for spike inference

To train the deep networks, the mean squared error between the smoothed ground truth spike rates and inferred spike rates was used as the loss function. This loss function not only optimizes the similarity of both signals (correlation), but also the absolute magnitude of the inferred spike rates. Based on errors computed via backpropagation, gradient descent was performed using a standard optimizer (*adagrad*; cf. Fig. S7). Based on a given resampled ground truth dataset, the network was trained using every single data point from this set, completing an epoch. Typically, training lasted for 10 epochs (except when analyzing overfitting in Fig. S7 and Fig. S8).

Crucially, in all spike inferences presented here, without any exception, a leave-one-out strategy was employed. For example, to infer the spike rates of a given neuron in a dataset, the network was trained on all neurons of this dataset except the neuron of interest. To infer spike rates for a given set of datasets, the training set always excluded the dataset for which inferences were made. This strategy of cross-validation is absolutely crucial and is strictly distinct from the process of fitting parameters for a neuron, which would yield better result for a given neuron but would fail to generalize to new data.

Architecture of alternative deep learning networks

All deep learning architectures (Fig. S8) were trained with the same loss function, the same input and the same optimizer as the default network.

Small convolutional filters network. The same architecture as the default network, with the only difference that smaller convolutional filter sizes were used, (15, 9, 3) instead of (31, 19, 5). Total of 9'891 trainable parameters.

Single convolutional layer network. Consisting of the first convolutional layer of the default network, a single max pooling layer and a single dense layer of 10 neurons. Total of 1'021 trainable parameters.

Deeper convolutional network (5 CNN layers). Consisting of 5 convolutional layers with filter sizes (11, 9, 7, 5, 3) and filter numbers (20, 30, 40, 40, 40), with three max pooling layers after the

second, fourth and fifth convolutional layers, and a final dense layer expansion of 10 neurons. The reduction of the filter sizes compared with the default network is necessary since no zero-padding was applied, resulting in a decrease of the size of the 1D trace with increasing layer depth. Total of 27'421 trainable parameters.

Deeper convolutional network (7 CNN layers). Consisting of 7 convolutional layers with filter sizes (7, 6, 5, 4, 3, 3, 3) and filter numbers (20, 30, 40, 40, 40, 40, 40), with three max pooling layers after the second, fifth and seventh convolutional layers, and a final dense layer expansion of 10 neurons. Total of 31'221 trainable parameters.

Batch normalization. Same as the default network, but with batch normalization[73] for regularization after each convolutional and dense layer, but before the respective ReLU transfer functions of each network layer. Total of 18'741 trainable parameters.

Locally connected network. Same as the default network, but with locally connected filters instead of convolutional filters. For convolutional filters, filter weights are shared across each position in the image space (here, in the temporal window), while the filters are different for each position for locally connected networks. The rationale behind this architecture is that different filters can be learned for each position, which is intuitive given that spike detection is not invariant to the position of the calcium transient in the window. Using different weights for each position of the filter sets results in a total of 229'231 trainable parameters.

Naïve LSTM model. LSTM units are complex neuronal units with internal states and gates that are used in recurrent networks to overcome the problem of vanishing gradients when backpropagating through time[74,75]. The time points of the input window are sequentially fed into the recurrent network, which are processed by the recurrent network, with earlier time points retained through recurrent activity or LSTM states and used to activate the network for processing of later time points. The investigated model consisted of two layers of each 25 LSTM units with ReLU as activation functions, followed by a simple dense expansion layer of 50 neurons with ReLU activation functions. Total of 4'051 trainable parameters.

Bi-directional LSTM model. The time points of the input window (64 data points) are split into past (32 data points) and future (32 data points) with respect to the time point used for spike inference ("presence"). Past and a reversed version of the future are each fed into a recurrent network based on a single layer of 25 LSTM units (with ReLU activations), such that the time point closest to "presence" is fed in last[76,77]. The output of the two recurrent networks for past and future is concatenated and connected with a dense fully connected layer of 50 simple units (ReLU activations). Total of 8'001 trainable parameters.

Linear network. Same as the default network, but with linear activation functions instead of rectifying linear units (ReLUs). The network is therefore entirely linear, but is based on the same architecture in terms of connectivity. Total of 18'541 trainable parameters.

Discretization of spiking probabilities

To obtain discrete spiking events from inferred probabilities, a brute-force fitting procedure was performed. The Gaussian kernel used to smooth the ground truth was used as a prior for the inferred spike rate that corresponds to a single action potential. The fit therefore consisted of optimally fitting a set of Gaussian kernels of the expected width and height to the inferred spike rate. We applied a brute-force approach to this problem, using a first guess that was then optimized by random modifications. The first guess was generated using Monte Carlo importance sampling, such that the overall number of discrete spikes matched the integral of inferred probabilities. Next, events were ranked in how they contributed to the fit by comparing the fit quality when single events were omitted. Lowest-ranking events were discarded and replaced by newly drawn events, again using importance sampling based on the residual probability distribution. Finally, each spike was shifted randomly over the entire duration and the best fit was used. Although this approach is relatively slow, it results in a reliable fit. To speed up the procedure, spiking probabilities were divided in continuous sequences of non-zero support (divide-and-conquer strategy). For the purpose of Fig. S11 and to allow for comparison against raw inferred spike rates, the resulting discrete spikes were convolved with the Gaussian smoothing kernel that had been used to generate the ground truth.

Generalized linear model to fit predictability across datasets

To predict how well a model trained on a given ground truth dataset (*e.g.*, DS#07) is able to infer activity for another dataset (*e.g.*, DS#13), a set of descriptors (regressors) was extracted for each dataset, and a generalized linear model (GLM) was trained to predict this relationship based on the regressors of the two respective datasets (Fig. S13). In total, 8 predictors were used, separately or together.

First, *indicator species* was set to 1 if training and test dataset had the same indicator species (synthetic dyes vs. genetically encoded dyes) and 0 otherwise. *Animal species* was set to 1 if training and test dataset had the same animal species (zebrafish vs. mouse) and 0 otherwise. *Spike rate* was computed as the absolute difference between median spike rates across neurons from training and test datasets. *Burstiness* was computed as the number of spikes that were spike within 50 ms of the timing of a given spike. This metric does quantifies the likelihood that a given spike is surrounded by other spikes. The *Fano factor* was computed by dividing the variance of inter-spike-intervals (ISIs) by the mean of ISIs[78]. Measured Fano factors were broadly distributed across datasets with a median of 3.7 and a standard deviation of 5.9, and an outlier dataset DS#17 in mouse CA3 with a Fano factor of 30.0. The *area of the linear kernel* was computed by summing up the area under the curve for the extracted linear kernel for each dataset. The *kernel decay constant* was computed without exponential fit by measuring the time between rise and decay time of the kernel directly. Rise and decay time points were identified by finding the first and last time point where the kernel surpassed 1/e of its maximum amplitude. The *correlation timecourse* was computed as the correlation of between the kernels of training and test dataset.

The GLM was fitted based on these regressors using the *glmfit()* command in Matlab with an identity linker function.

Adaptation of model-based spike inference algorithms

The MLSpike algorithm was downloaded from https://github.com/MLspike/spikes and used within Matlab 2017a [11]. Parameter settings were manually explored for several datasets using the graphical demo user interface. Then, some parameters (noise level *sigma* and inverse frame rate *dt*) were fixed to the values constrained by the ground truth. The *drift* parameter was set to 0.1. For synthetic dyes (DS#01-04), a saturating non-linearity (*saturation* = 0.01) was used, whereas for all other datasets a GCaMP-like nonlinearity (*pnonlin* = [1.0 0.0]) was defined and kept the same across datasets, since predictions have been described to only slightly depend on the precise values of the non-linearity[11]. Based on manual exploration, the two parameters *tau* (decay time constant) and *amplitude* (amplitude of a single action potential) were explored in a grid search for all ground truth datasets separately. The grid search ranged from 0.1 to 5 s for *tau* and from 0.01 to 0.35 for *amplitude*.

The Peeling algorithm was downloaded from https://github.com/HelmchenLab/CalciumSim and used within Matlab 2017a [10]. A single-exponential linear model with default values was used. A grid search was performed over two parameters for all ground truth datasets: time constant of the exponential decay (*tau1*) and the amplitude of a single spike (*amp1*). Grid search ranged from 0.25 – 5 s for *tau1* and from 2.5 – 35 for *amp1*.

The Python implementation of the OASIS algorithm was downloaded from https://github.com/j-friedrich/OASIS and used within Python 3.7 [14]. The constrained version of OASIS was used to reduce the number of free parameters, with only one single free parameter, *g*. *g* relates to the exponential time fluorescence decay constant τ with the frame rate f via the formula: $g = e^{-1/\tau f}$. Grid search was performed for *g* in the range between 0.02 and 0.98, with a granularity of 0.02.

The optimal parameters as results of the grid searches are listed in Table 2.

Computational cost of spike inference

The four investigated algorithms exhibit different properties when scaling up the length of the calcium traces. For example, MLSpike and Peeling suffer from supra-linear cost when the duration of an analyzed calcium trace is increased, while CASCADE shows the opposite behavior due to its capability to parallelize spike inference. Therefore, all 20 full ground truth datasets, resampled at a noise level of 2 and a frame rate of 7.5 Hz, were used as a benchmark, consisting of recordings ranging from 10s of seconds up to several minutes. Processing time was averaged across all data points from all datasets. The time required to load the data from hard disk was not included. For CASCADE, the time for pre-processing the raw calcium data in order to generate a 64 point-wide for each time point was included in the benchmarking.

Unsupervised sequence extraction using seqNMF

The Matlab-based toolbox seqNMF was used to extract temporal patterns for Fig. 6 in an unsupervised fashion[43]. Based on initial parameter exploration, we used the following settings: K=7, L=20 and λ=0.002. K indicates the number of extracted patterns, L the number of time points for each pattern, λ serves as a regularizer to decorrelate the detected patterns[43]. The result of this unsupervised non-negative matrix factorization approach are K=7 temporal patterns that are each of them associated with a temporal loading which indicates when the temporal pattern became active. The temporal patterns and the temporal loadings provide low-complexity factors that break down the more complex population dynamics (Fig. 6).

Allen Brain Observatory data

The complete calcium imaging data of the Allen Brain Observatory Visual Coding dataset were downloaded from http://observatory.brain-map.org/visualcoding via the AllenSDK with a Python interface. Layers were assigned based on imaging depth as described[44]. Imaging depth, transgenic lines, cortical areas and fluorescence traces were extracted from NWB files. For analysis, neuropil-corrected calcium traces from the Allen Brain Observatory dataset were used. Since all recordings were performed at an imaging rate of approximately 30 Hz, a single set of CASCADE models (‚universal model' at 30 Hz, Fig. 4a) was used to predict spiking activity.

Statistical tests and box plots

Statistical analysis was performed in Matlab 2017a and R. Only non-parametric tests were used. The Mann-Whitney rank sum test was used for non-paired samples (*e.g.*, comparison across datasets) and the Wilcoxon signed-rank test for paired samples (*e.g.*, comparison of predictions for the same set of neurons using two different algorithms). Mostly, two-sided tests were applied except when otherwise indicated. Effect sizes Δ±CI (pseudo-median and 95% confidence intervals unless otherwise indicated) were computed in R. Boxplots used standard settings in Matlab, with the central line at the median of the distribution, the box at the 25th and 75th percentile and the whiskers at the location of approximately 99.3 percent coverage for the case of a normal distribution.

REFERENCES

1. Göbel, W. & Helmchen, F. In Vivo Calcium Imaging of Neural Network Function. *Physiology* **22**, 358–365 (2007).
2. Harris, K. D., Quiroga, R. Q., Freeman, J. & Smith, S. L. Improving data quality in neuronal population recordings. *Nat. Neurosci.* **19**, 1165–1174 (2016).
3. Rose, T., Goltstein, P. M., Portugues, R. & Griesbeck, O. Putting a finishing touch on GECIs. *Front. Mol. Neurosci.* **7**, (2014).
4. Sabatini, B. L. The impact of reporter kinetics on the interpretation of data gathered with fluorescent reporters. *bioRxiv* 834895 (2019).
5. Wei, Z. *et al.* A comparison of neuronal population dynamics measured with calcium imaging and electrophysiology. *bioRxiv* 840686 (2019).
6. Yaksi, E. & Friedrich, R. W. Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca2+ imaging. *Nat. Methods* **3**, 377–383 (2006).
7. Greenberg, D. S., Houweling, A. R. & Kerr, J. N. D. Population imaging of ongoing neuronal activity in the visual cortex of awake rats. *Nat. Neurosci.* **11**, 749–751 (2008).
8. Vogelstein, J. T. *et al.* Spike inference from calcium imaging using sequential Monte Carlo methods. *Biophys. J.* **97**, 636–655 (2009).
9. Vogelstein, J. T. *et al.* Fast Nonnegative Deconvolution for Spike Train Inference From Population Calcium Imaging. *J. Neurophysiol.* **104**, 3691–3704 (2010).
10. Lütcke, H., Gerhard, F., Zenke, F., Gerstner, W. & Helmchen, F. Inference of neuronal network spike dynamics and topology from calcium imaging data. *Front. Neural Circuits* **7**, (2013).
11. Deneux, T. *et al.* Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo. *Nat. Commun.* **7**, 1–17 (2016).
12. Greenberg, D. S. *et al.* Accurate action potential inference from a calcium sensor protein through biophysical modeling. *bioRxiv* 479055 (2018).
13. Pachitariu, M., Stringer, C. & Harris, K. D. Robustness of Spike Deconvolution for Neuronal Calcium Imaging. *J. Neurosci.* **38**, 7976–7985 (2018).
14. Friedrich, J., Zhou, P. & Paninski, L. Fast online deconvolution of calcium imaging data. *PLOS Comput. Biol.* **13**, e1005423 (2017).
15. Berens, P. *et al.* Community-based benchmarking improves spike rate inference from two-photon calcium imaging data. *PLoS Comput. Biol.* **14**, (2018).
16. Jewell, S. & Witten, D. Exact spike inference via L0 optimization. *Ann. Appl. Stat.* **12**, 2457–2482 (2018).
17. Sasaki, T., Takahashi, N., Matsuki, N. & Ikegaya, Y. Fast and Accurate Detection of Action Potentials From Somatic Calcium Fluctuations. *J. Neurophysiol.* **100**, 1668–1676 (2008).
18. Theis, L. *et al.* Benchmarking Spike Rate Inference in Population Calcium Imaging. *Neuron* **90**, 471–482 (2016).
19. Sebastian, J., Sur, M., Murthy, H. A. & Magimai.-Doss, M. Signal-to-signal networks for improved spike estimation from calcium imaging data. *bioRxiv* 2020.05.01.071993 (2020).
20. Éltes, T., Szoboszlay, M., Kerti-Szigeti, K. & Nusser, Z. Improved spike inference accuracy by estimating the peak amplitude of unitary [Ca2+] transients in weakly GCaMP6f-expressing hippocampal pyramidal cells. *J. Physiol.* **597**, 2925–2947 (2019).
21. Evans, M. H., Petersen, R. S. & Humphries, M. D. On the use of calcium deconvolution algorithms in practical contexts. *bioRxiv* 871137 (2019).
22. Zhu, P., Fajardo, O., Shum, J., Zhang Schärer, Y.-P. & Friedrich, R. W. High-resolution optical control of spatiotemporal neuronal activity patterns in zebrafish using a digital micromirror device. *Nat. Protoc.* **7**, 1410–1425 (2012).
23. Bethge, P. *et al.* An R-CaMP1.07 reporter mouse for cell-type-specific expression of a sensitive red fluorescent calcium indicator. *PLOS ONE* **12**, e0179460 (2017).
24. Tada, M., Takeuchi, A., Hashizume, M., Kitamura, K. & Kano, M. A highly sensitive fluorescent indicator dye for calcium imaging of neural activity in vitro and in vivo. *Eur. J. Neurosci.* **39**, 1720–1728 (2014).

25. Huang, L. *et al.* Relationship between spiking activity and simultaneously recorded fluorescence signals in transgenic mice expressing GCaMP6. *bioRxiv* 788802 (2019).
26. Ledochowitsch, P. *et al.* On the correspondence of electrical and optical physiology in in vivo population-scale two-photon calcium imaging. *bioRxiv* 800102 (2019).
27. Dana, H. *et al.* Sensitive red protein calcium indicators for imaging neural activity. *eLife* **5**, e12727 (2016).
28. Akerboom, J. *et al.* Optimization of a GCaMP calcium indicator for neural activity imaging. *J. Neurosci. Off. J. Soc. Neurosci.* **32**, 13819–13840 (2012).
29. Chen, T.-W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
30. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
31. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
32. Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M. & Tolias, A. S. Engineering a Less Artificial Intelligence. *Neuron* **103**, 967–979 (2019).
33. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
34. Nath, T. *et al.* Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat. Protoc.* **14**, 2152–2176 (2019).
35. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009).
36. Denis, J., Dard, R. F., Quiroli, E., Cossart, R. & Picardo, M. A. DeepCINAC: a deep-learning-based Python toolbox for inferring calcium imaging neuronal activity based on movie visualization. *bioRxiv* 803726 (2020).
37. Gauthier, J. L. *et al.* Detecting and Correcting False Transients in Calcium Imaging. *bioRxiv* 473470 (2018).
38. Giovannucci, A. *et al.* CaImAn an open source tool for scalable calcium imaging data analysis. *eLife* **8**, e38173 (2019).
39. Keemink, S. W. *et al.* FISSA: A neuropil decontamination toolbox for calcium imaging signals. *Sci. Rep.* **8**, 3493 (2018).
40. Rupprecht, P., Prendergast, A., Wyart, C. & Friedrich, R. W. Remote z-scanning with a macroscopic voice coil motor for fast 3D multiphoton laser scanning microscopy. *Biomed. Opt. Express* **7**, 1656–1671 (2016).
41. Blumhagen, F. *et al.* Neuronal filtering of multiplexed odour representations. *Nature* **479**, 493–498 (2011).
42. Rupprecht, P. & Friedrich, R. W. Precise Synaptic Balance in the Zebrafish Homolog of Olfactory Cortex. *Neuron* **100**, 669-683.e5 (2018).
43. Mackevicius, E. L. *et al.* Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience. *eLife* **8**, e38471 (2019).
44. de Vries, S. E. J. *et al.* A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nat. Neurosci.* **23**, 138–151 (2020).
45. Lin, I.-C., Okun, M., Carandini, M. & Harris, K. D. The Nature of Shared Cortical Variability. *Neuron* **87**, 644–656 (2015).
46. Moreno-Bote, R. *et al.* Information-limiting correlations. *Nat. Neurosci.* **17**, 1410–1417 (2014).
47. Williams, A. H. *et al.* Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron* **98**, 1099-1115.e8 (2018).
48. Pachitariu, M. *et al.* Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *bioRxiv* 061507 (2017).
49. Kaifosh, P., Zaremba, J. D., Danielson, N. B. & Losonczy, A. SIMA: Python software for analysis of dynamic fluorescence imaging data. *Front. Neuroinformatics* **8**, (2014).
50. Charles, A. S., Song, A., Gauthier, J. L., Pillow, J. W. & Tank, D. W. Neural Anatomy and Optical Microscopy (NAOMi) Simulation for evaluating calcium imaging methods. *bioRxiv* 726174 (2019).
51. Siegle, J. H. *et al.* Reconciling functional differences in populations of neurons recorded with two-photon imaging and electrophysiology. *bioRxiv* 2020.08.10.244723 (2020).
52. Kay, K. *et al.* Constant Sub-second Cycling between Representations of Possible Futures in the Hippocampus. *Cell* **180**, 552-567.e25 (2020).

53. Bourg, A. van der *et al.* Temporal refinement of sensory-evoked activity across layers in developing mouse barrel cortex. *Eur. J. Neurosci.* **50**, 2955–2969 (2019).
54. Chen, I.-W., Papagiakoumou, E. & Emiliani, V. Towards circuit optogenetics. *Curr. Opin. Neurobiol.* **50**, 179–189 (2018).
55. Packer, A. M., Russell, L. E., Dalgleish, H. W. P. & Häusser, M. Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo. *Nat. Methods* **12**, 140–146 (2015).
56. Pégard, N. C. *et al.* Three-dimensional scanless holographic optogenetics with temporal focusing (3D-SHOT). *Nat. Commun.* **8**, 1228 (2017).
57. Griffiths, V. A. *et al.* Real-time 3D movement correction for two-photon imaging in behaving animals. *Nat. Methods* **17**, 741–748 (2020).
58. Inoue, M. *et al.* Rational Engineering of XCaMPs, a Multicolor GECI Suite for In Vivo Imaging of Complex Brain Circuit Dynamics. *Cell* **177**, 1346-1360.e24 (2019).
59. Khan, A. G. *et al.* Distinct learning-induced changes in stimulus selectivity and interactions of GABAergic interneuron classes in visual cortex. *Nat. Neurosci.* **21**, 851–859 (2018).
60. Fort, S., Hu, H. & Lakshminarayanan, B. Deep Ensembles: A Loss Landscape Perspective. *ArXiv191202757 Cs Stat* (2019).
61. 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).
62. Frank, T., Mönig, N. R., Satou, C., Higashijima, S. & Friedrich, R. W. Associative conditioning remaps odor representations and modifies inhibition in a higher olfactory brain area. *Nat. Neurosci.* **22**, 1844–1856 (2019).
63. Kitamura, K., Judkewitz, B., Kano, M., Denk, W. & Häusser, M. Targeted patch-clamp recordings and single-cell electroporation of unlabeled neurons in vivo. *Nat. Methods* **5**, 61–67 (2008).
64. Perkins, K. L. Cell-attached voltage-clamp and current-clamp recording and stimulation techniques in brain slices. *J. Neurosci. Methods* **154**, 1–18 (2006).
65. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: Flexible software for operating laser scanning microscopes. *Biomed. Eng. OnLine* **2**, 13 (2003).
66. Suter, B. A. *et al.* Ephus: Multipurpose Data Acquisition Software for Neuroscience Experiments. *Front. Neural Circuits* **4**, (2010).
67. Huang, K.-H., Rupprecht, P., Frank, T., Kawakami, K. & Friedrich, R. A virtual reality system to analyze neural activity and behavior in adult zebrafish. *Nat. Methods* **17**(3), 343-351 (2020).
68. Pernía-Andrade, A. J. *et al.* A Deconvolution-Based Method with High Sensitivity and Temporal Resolution for Detection of Spontaneous Synaptic Currents In Vitro and In Vivo. *Biophys. J.* **103**, 1429–1439 (2012).
69. Guzman, S. J., Schlögl, A. & Schmidt-Hieber, C. Stimfit: quantifying electrophysiological data with Python. *Front. Neuroinformatics* **8**, (2014).
70. GENIE project, J. F. C., HHMI & Svoboda, K. Simultaneous imaging and loose-seal cell-attached electrical recordings from neurons expressing a variety of genetically encoded calcium indicators. *CRCNS.org.* (2015).
71. Boaz, M., Dana, H., Kim, D. S., Svoboda, K. & GENIE project, J. F. C., HHMI. jRGECO1a and jRCaMP1a characterization in the intact mouse visual cortex, using AAV-based gene transfer, 2-photon imaging and loose-seal cell attached recordings, as described in Dana et al 2016. *CRCNS.org.* (2016).
72. Reynolds, S., Abrahamsson, T., Sjöström, P. J., Schultz, S. R. & Dragotti, P. L. CosMIC: A Consistent Metric for Spike Inference from Calcium Imaging. *Neural Comput.* **30**, 2726–2756 (2018).
73. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv 150203167 Cs* (2015).
74. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).
75. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: continual prediction with LSTM. 850–855 (1999).
76. Schuster, M. & Paliwal, K. Bidirectional recurrent neural networks. *Signal Process. IEEE Trans.* **45**, 2673–2681 (1997).
77. Graves, A., Fernandez, S. & Schmidhuber, J. Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. *International Conference on Artificial Neural Networks.* Springer, Berlin, Heidelberg, 2005.
78. Eden, U. T. & Kramer, M. A. Drawing inferences from Fano factor calculations. *J. Neurosci. Methods* **190**, 149–152 (2010).

Supplementary information for "A deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging data"

Peter Rupprecht†, Stefano Carta, Adrian Hoffmann, Mayumi Echizen, Kazuo Kitamura, Fritjof Helmchen†, Rainer W. Friedrich†

† Corresponding authors: rupprecht@hifo.uzh.ch, helmchen@hifo.uzh.ch, rainer.friedrich@fmi.ch

- 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 S1-S19
- Supplementary Table (Table 2)

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. S4).

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. S4a). 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. S4a). 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 $n = 500$. 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. S4a,b). 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: $\Delta_1 = 0.010 \pm 0.004$, pseudo-median ± 95% C.I., $p < 1e\text{-}4$, paired Wilcoxon test; decrease for Poisson noise: $\Delta_1 = 0.006 \pm 0.004$, $p < 0.005$). 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: $\Delta_2 = 0.006 \pm 0.008$, $p = 0.13$; Poisson noise: $\Delta_2 = 0.005 \pm 0.009$, $p = 0.28$). 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 (Fig. S4b). 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. S7a-e). All of these observations could be confirmed over a surprisingly large range, indicating that the network performance is highly robust with respect to any hyper-parameter choices.

In addition, we also investigated potential overfitting of the training dataset. We observed that 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. S7f). This learning behavior suggests only moderate overfitting. At the same time the training loss decreased monotonically (Fig. S7g). 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. S7f), individual network instances sometimes reached unfavorable states. As expected from known properties of deep networks[60], this effect could be easily eliminated by ensemble averaging over 5 networks (Fig. S7h).

Finally, we also tested the performance when employing a network architecture different from the standard convolutional architecture (hence called '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. S8a). 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. S8b), consistent with previous observations[15].

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. S8a,b), which prevents the algorithm from non-linearly adjusting decision boundaries.

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[10,11,14]. However, the precise identification of individual spikes with high temporal resolution is difficult due to shot noise, other noise sources, low sampling rates and the non-linear response of calcium indicators[3,25,26,61]. It is therefore not clear whether discrete spikes can be reliably inferred. 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. S11a-d). However, this procedure can also enhance small false positive or small negative errors (red arrows in Fig. S11a-d). Over all ground truth datasets, the correlation metric slightly but consistently decreased when spike rates were discretized (Fig. S11e), indicating that the data quality did not allow for efficient use of the prior. Although discretized spike rates tended to decrease the error (Fig. S11e), 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. S11d,f). Together, this suggests that the available datasets do not allow to discretize 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 demo script in the public repository.

SUPPLEMENTARY FIGURES

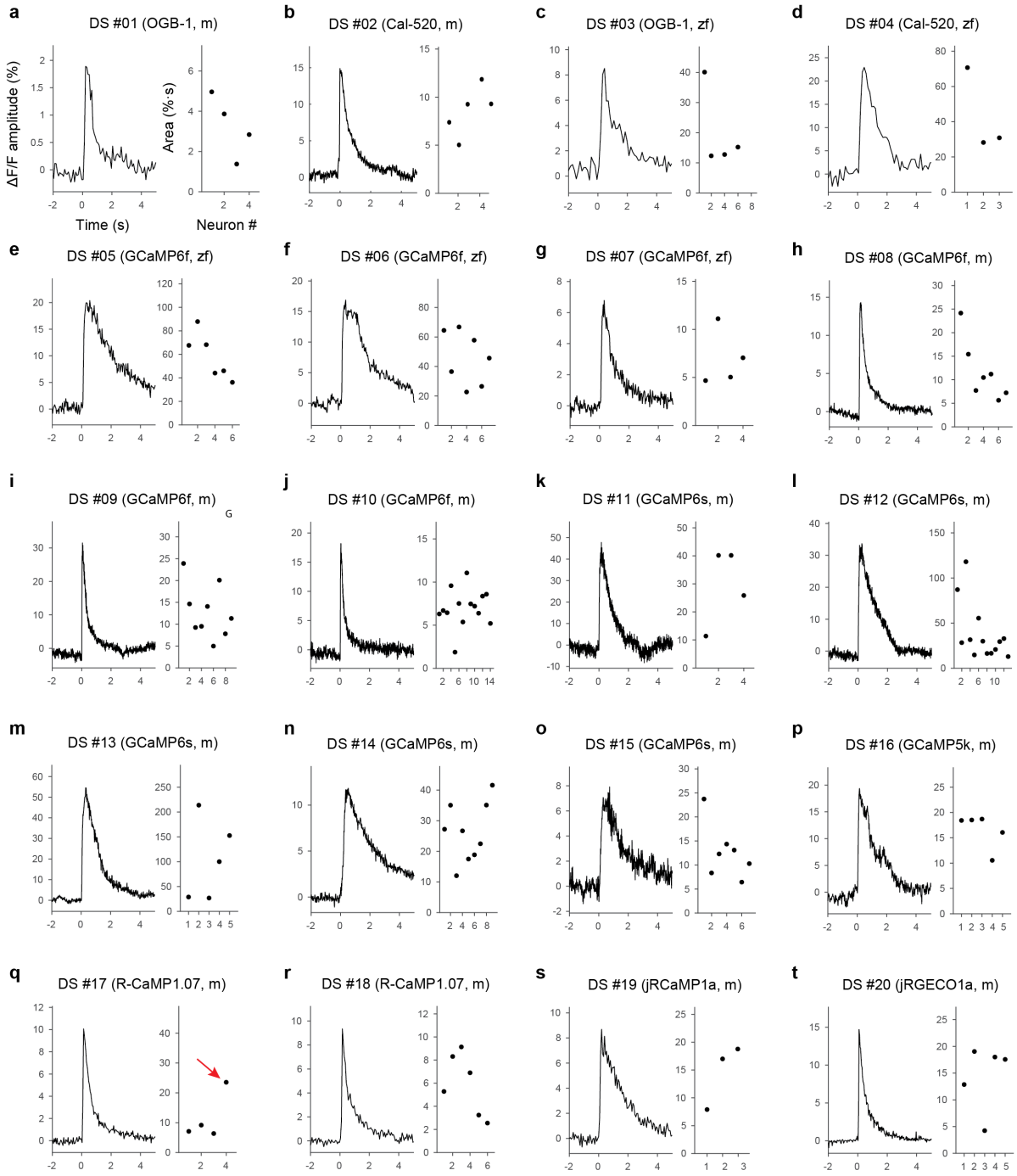

Figure S1 | Linear kernels extracted from all 20 ground truth datasets. The kernels are optimized such that when the ground truth spike times are linearly convolved with the kernel, the experimentally recorded ΔF/F trace is ideally approximated. In practice, this is achieved using regularized linear deconvolution of calcium traces based on spike times (Methods). Kernels vary both in amplitude and shape across datasets and within datasets. For single neurons, the kernel area (right panels) are only shown if the kernel could be reliably determined, as tested with the variability of the kernel across the recording (Methods). The red arrow in panel (q) indicates an outlier case, which is discussed in Fig. 3a. m: Mouse, zf: Zebrafish.

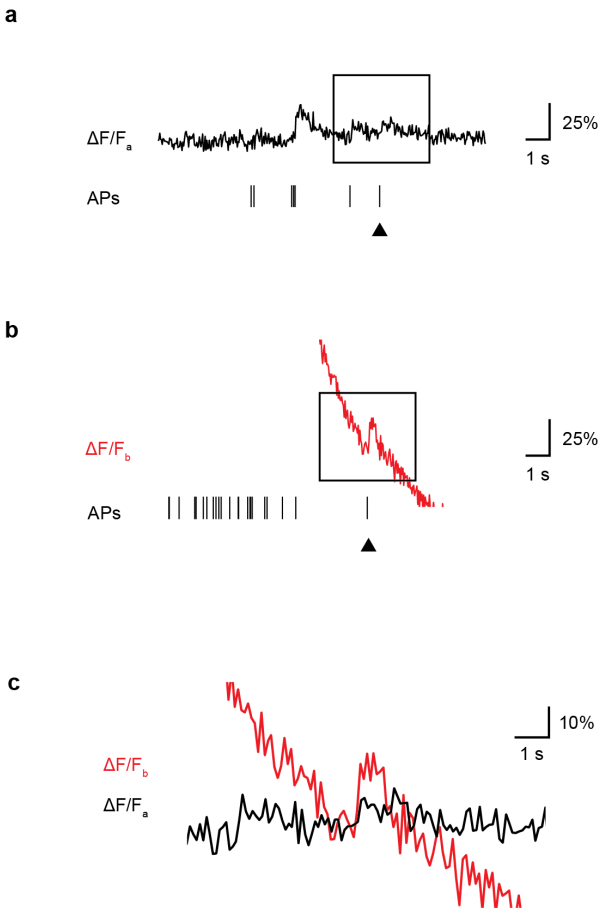

Figure S2 | Example of history-dependence of spike-evoked fluorescence. Calcium recordings of the same neuron during a phase of sparse spiking **(a)** and after a burst **(b)** are shown. Single spikes are marked with a black arrow head and overlayed and magnified in **(c)**, clearly showing a much larger 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#05.

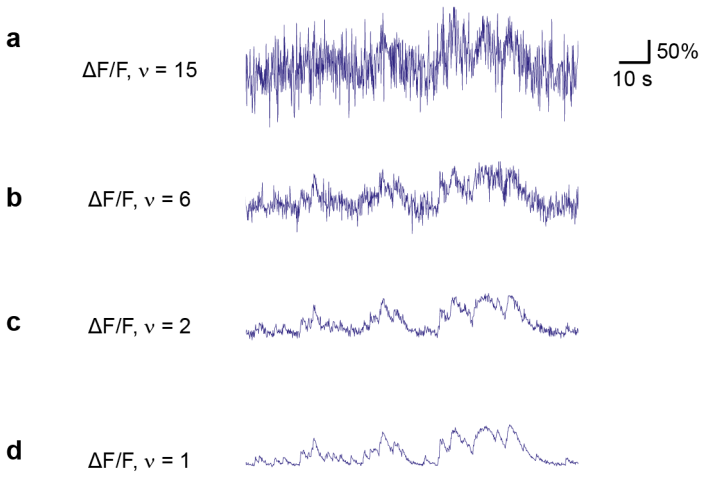

Figure S3 | Illustration of different baseline noise levels. ΔF/F ground truth traces were resampled with added noise to reach the target noise level ν. **a-d**, Noise level illustration from $\nu = 15\% \cdot s^{-1/2}$ (very high noise level) to $\nu = 1\% \cdot s^{-1/2}$ (very low noise level).

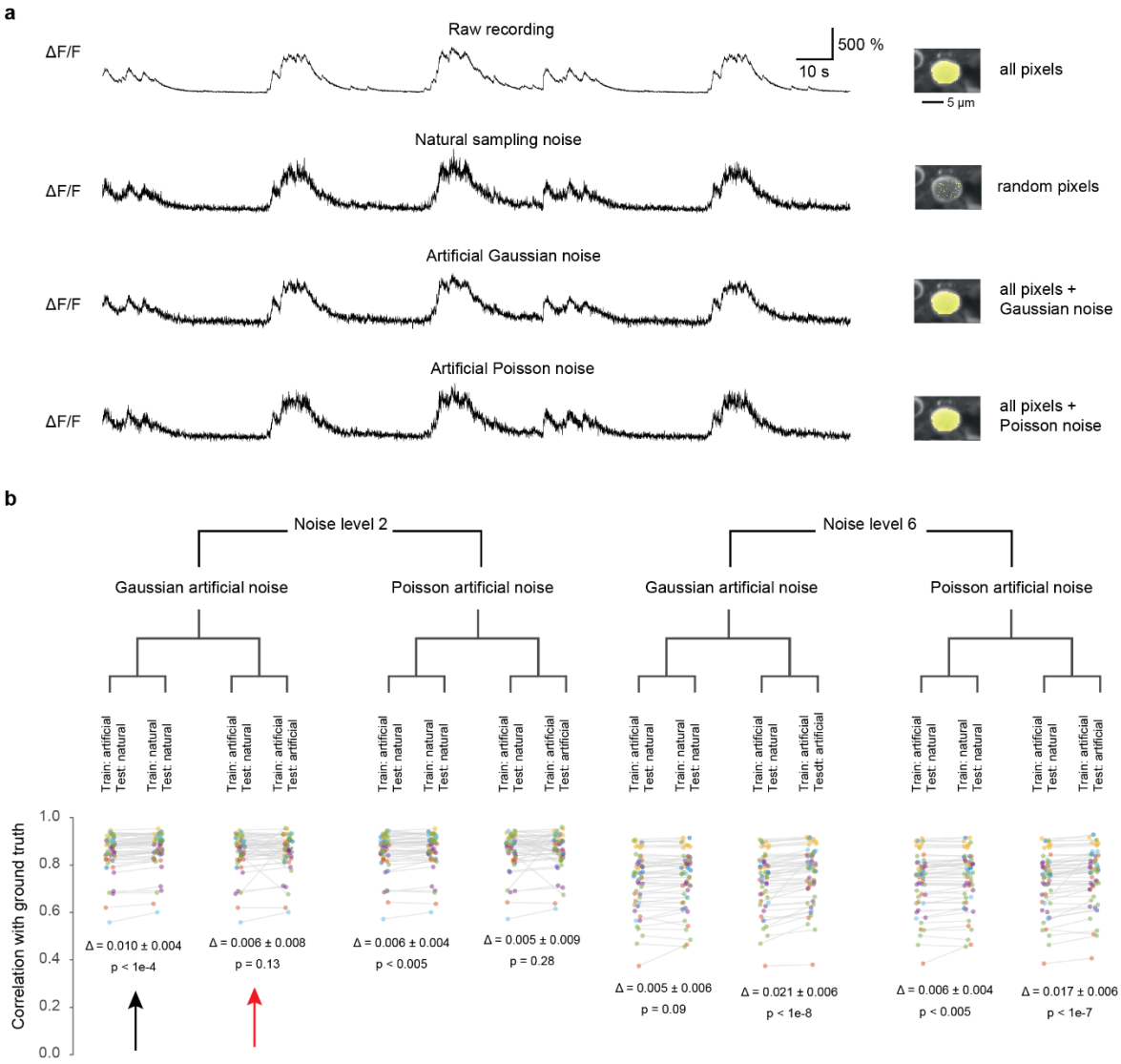

Figure S4 | Natural resampling of the ground truth vs. artificially sampled noise. a, Row 1: Raw ground truth calcium recording, mean fluorescence. Row 2: Same ground truth recording, spatial subsampling (natural sampling) of random pixels in the ROI. Row 3: Same recording, mean fluorescence with added Gaussian noise. Row 4: Same recording, mean fluorescence with added Poisson noise. **b,** Testing different noise modes for training and testing at two different noise levels (2 = left branch, 6 = right branch). For each condition (noise level; mode of artificial noise), two comparisons between models are made: 1) Trained with artificial and tested with natural noise vs. trained and tested with natural noise (example with black arrow). 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 are computed as pseudo-median ± 95% confidence intervals. P-values and pseudo-medians were computed using paired Wilcoxon signed-rank tests.

Figure S5. 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.

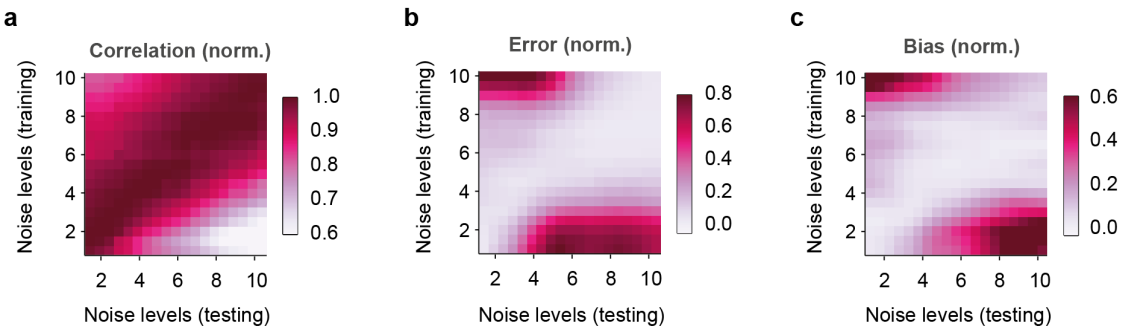

Figure S6 | **Matching noise-levels of training and test data.** Same as Fig. 2f-h, but with each column (testing level) normalized in order to highlight that the optimal training level for each testing noise level lies close to the diagonal. The correlation (a) was normalized by the maximum of each column, while error and bias metrics have been normalized by the minimum in this plot.

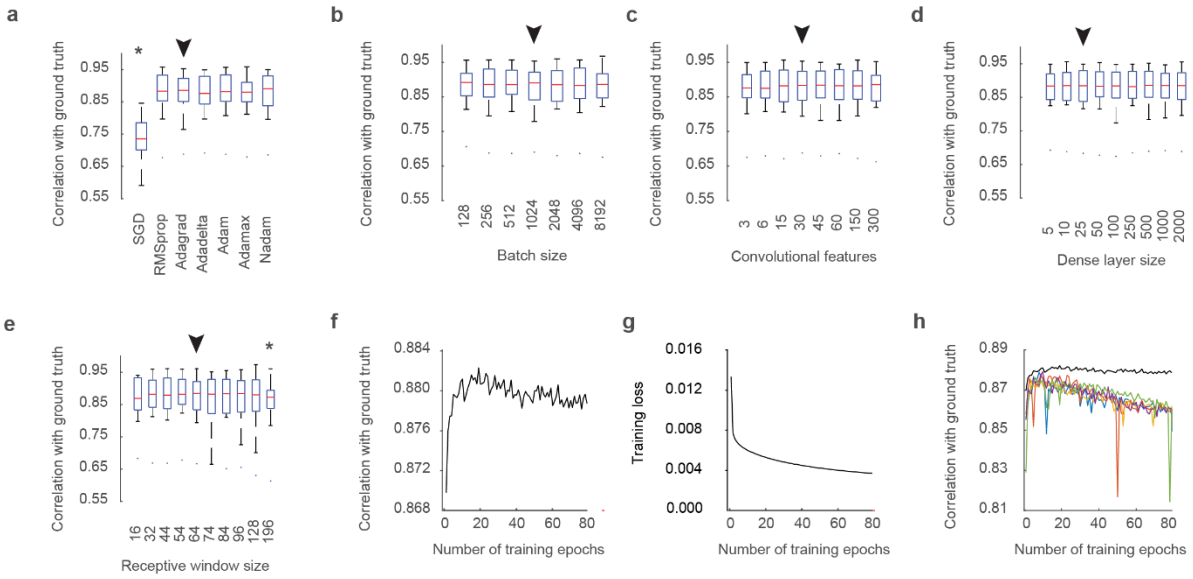

Figure S7 | Parameter-robustness of CASCADE with respect to hyper-parameters. a, Performance is robust with respect to choice of the gradient descent optimizer, unless the naive stochastic gradient descent (SGD) is selected. **b,** Performance is robust with respect to the batch size used during learning. 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; a receptive window of 64 data points corresponds *e.g.* to 64/7.5 ≈ 8.5 s in this setting. All parameter studies were performed in DS#03 at a calcium imaging rate of 7.5 Hz and a resampled noise level of 2. Networks were trained for 10 epochs with all data except for a single neuron, which was used for testing. Ensembles of 5 networks were used, and results were averaged across 3 iterations. No significant differences as observed by a Wilcoxon paired signed-rank test were found (exceptions indicated by asterisks). The parameter choices used as default values for the model are indicated by a black arrowhead. **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).

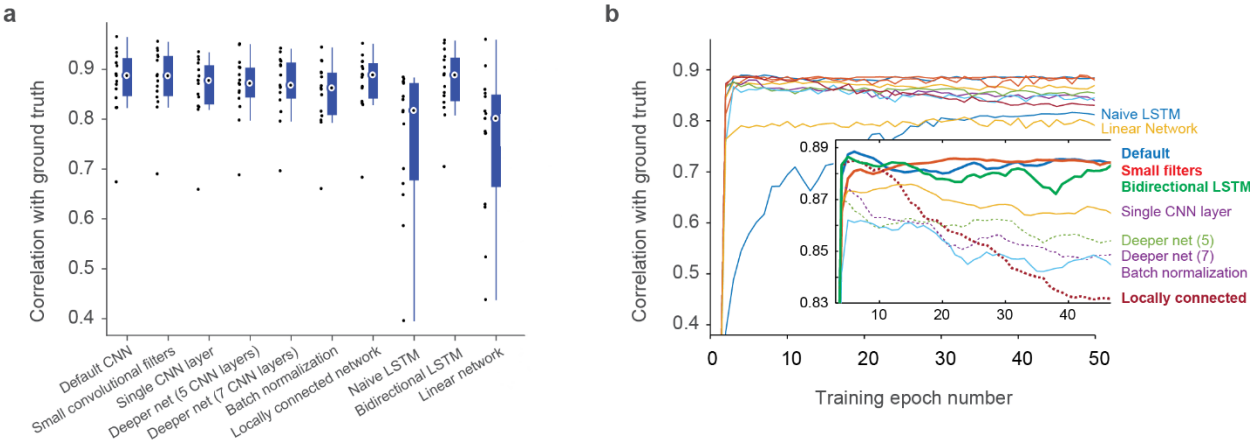

Figure S8 | **Comparison across deep learning architectures. a,** Comparison of the default network with a network with smaller filters (Δ = -(0.002 ± 0.006), p = 0.28; paired Wilcoxon signed-rank test; pseudo-median Δ ± 95% confidence interval), with a single layer network (Δ = 0.013 ± 0.009, p = 0.008), 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#03), sampled at 7.5 Hz at a noise level of 2. For detailed descriptions of all architectures see Methods. **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).

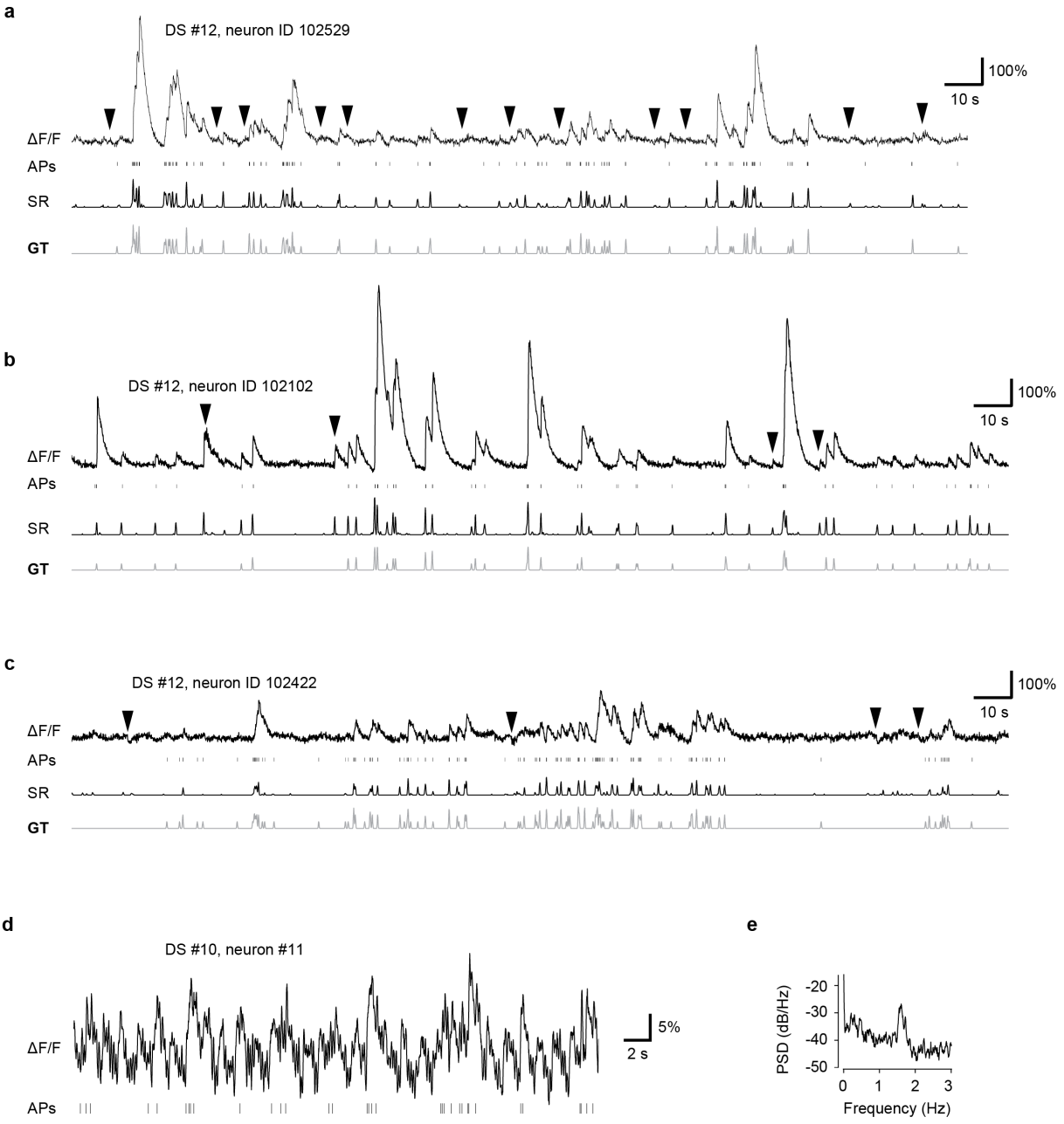

Figure S9 | Typical artifacts in ground truth recordings. Calcium trace (ΔF/F), true action potentials (APs), inferred spiking activity (SR) and true ground truth spiking activity (GT). **a,** The baseline of this recording is unstable, exhibiting irregular bumps (arrowheads). The supervised deep network can learn to ignore these movement artifacts if their dynamics is dissimilar from the sharp onset of calcium transients. Predictions of the deep network are shown in black, ground truth in grey. **b,** Fluorescence transients without corresponding action potentials are clearly visible (arrowheads). These are induced by contamination through bright neuropil. The deep network is unable to distinguish this artifact from true calcium transients. **c,** Negative transients (arrowheads) are generated by standard neuropil decontamination (subtraction of the neuropil surround). The deep network can learn to partially ignore these events. **d,** Trace showing periodic movement artifacts that do not correspond to action potentials. **e,** A power spectral density of the recording in (d) exhibits a peak at ca. 1.5 Hz, suggesting breathing of the anesthetized animal underlying the movement artifact.

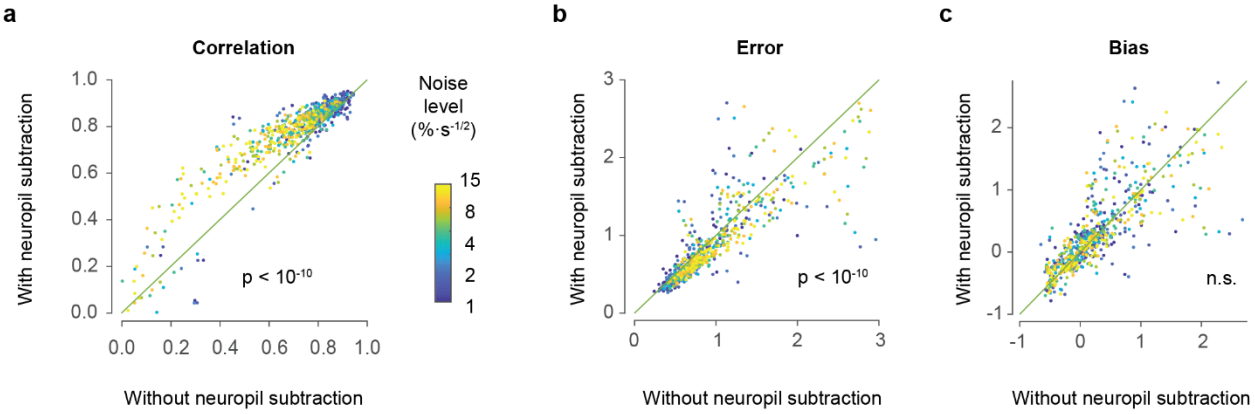

Figure S10 | **Neuropil decontamination improves spike inference.** For datasets DS#09-12 (Huang et al. (2019) datasets from the Allen Institute), ground truth datasets were extracted both with and without simple subtractive neuropil decontamination. **a,** The performance (correlation) is improved by neuropil decontamination. The same calcium recordings were analyzed for several noise levels (color-coded). The p-value (paired 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.

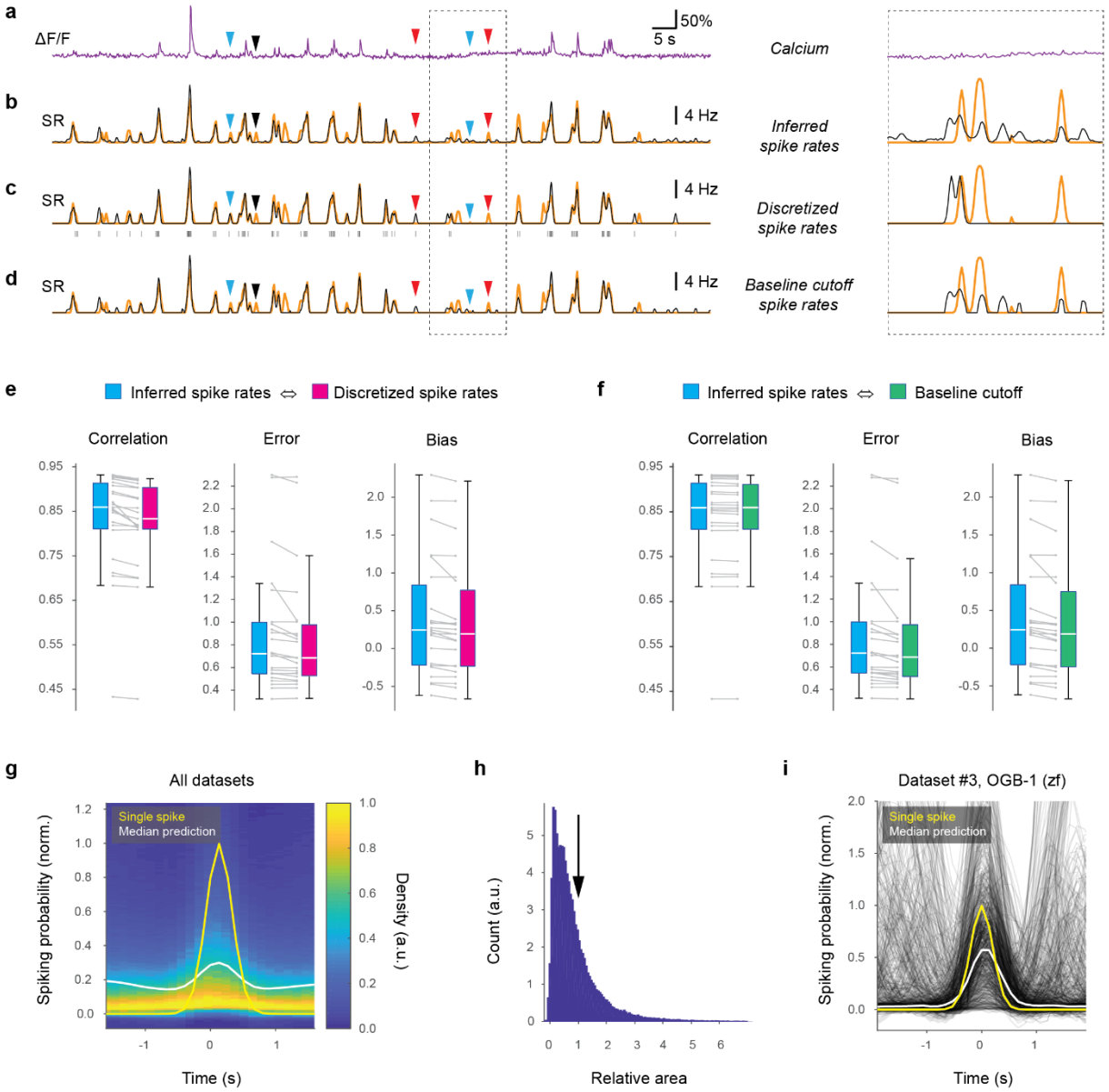

Figure S11 | Analysis of single-spike resolution of spike inference. a, Calcium $\Delta F/F_0$ of an example neuron. **b,** Inferred spike rate (SR, black) and ground truth spike rate (orange). Inset to the right zooms into a section of the recording to highlight the noise floor. **c,** Discrete spikes 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. **f,** Thresholding the predictions as in (d) results in the same reduction of error and biases, without negatively affecting the correlations. **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 (white) (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; 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#03; 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.

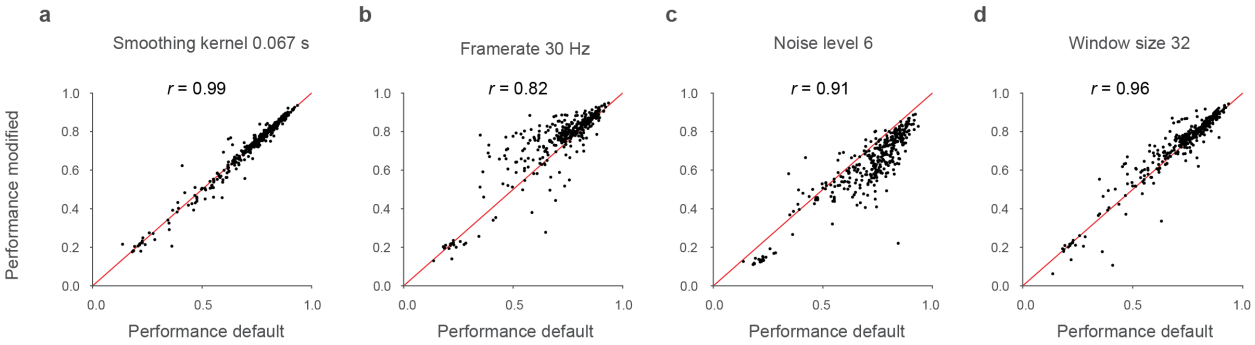

Figure S12 | **Parameter changes leave mutual predictability across datasets largely unchanged.** The off-diagonal elements in Fig. 4a were recomputed using different parameter settings. The standard parameters were a smoothing kernel of 0.2 s, a framerate of 7.5 Hz, a noise level of 2 and a window size 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. 4a. The red line indicates the unity function. Correlation between the two conditions is indicated as *r*. **a,** Smoothing kernel reduced to 0.067 s. **b,** Framerate increased to 30 Hz. **c,** Noise level increased to 6. **d,** Window size decreased to 32.

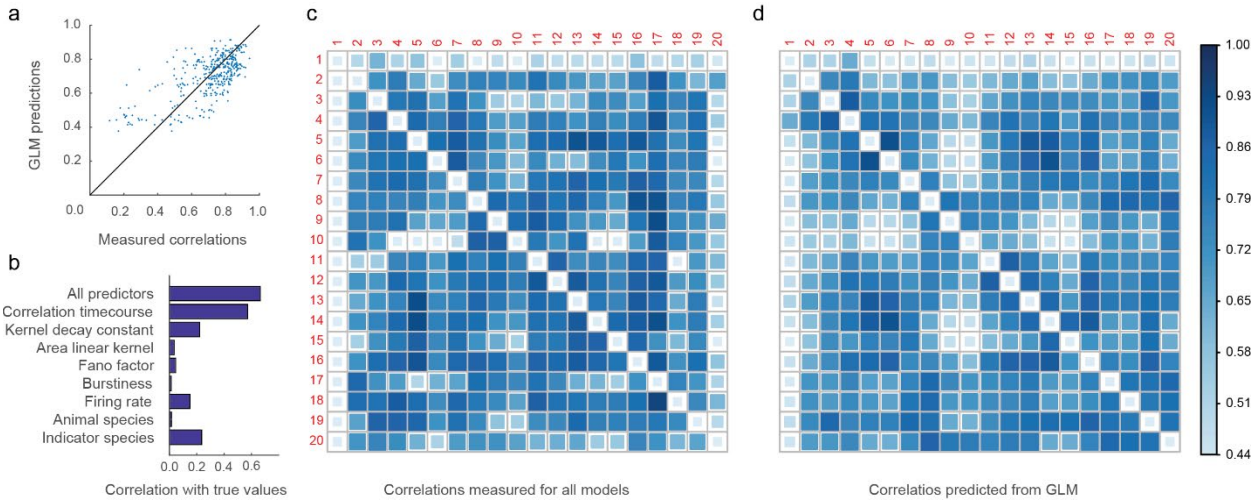

Figure S13 | Predicting cross-dataset predictability with a generalized linear model (GLM). We considered several characteristics of ground truth datasets and evaluated whether they could improve the mutual predictability across ground truth datasets. This includes characteristics that are accessible 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. Fig. S1; 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. 4a). **b,** A GLM based on “all predictors” results in a correlation that recovers panel (a). Using only one of the predictors reduces, often very significantly, the correlation. **c,** Measured cross-dataset predictability (reproducing Fig. 4a). **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 ‘universal’ model that was trained on all reliable datasets (Fig. 4a)

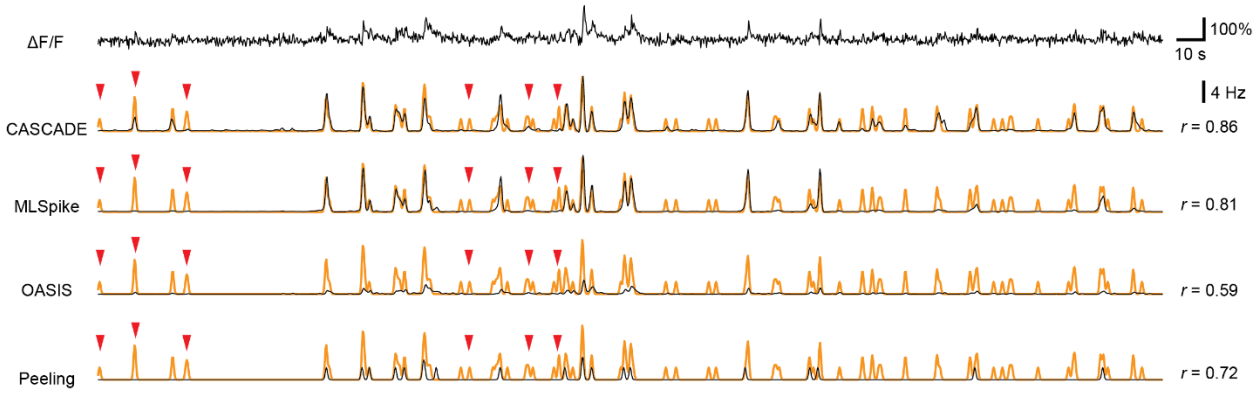

Figure S14 | Comparison with model-based algorithms, extension of Fig. 5a. Example predictions from the deep-learning based method (CASCADE) and three model-based algorithms (MLSpike, OASIS, Peeling) of a ΔF/F recording. Inferred spike rates are in black, ground truth spike rates in orange. *r* indicates correlation of predictions with ground truth. Events that are not detected across all algorithms (false negatives) are labeled with red arrowheads. Compared to the example in Fig. 5a, the calcium recording here is rather noisy due to movement-induced artifacts and the insensitivity of GCaMP to single action potentials in this neuron.

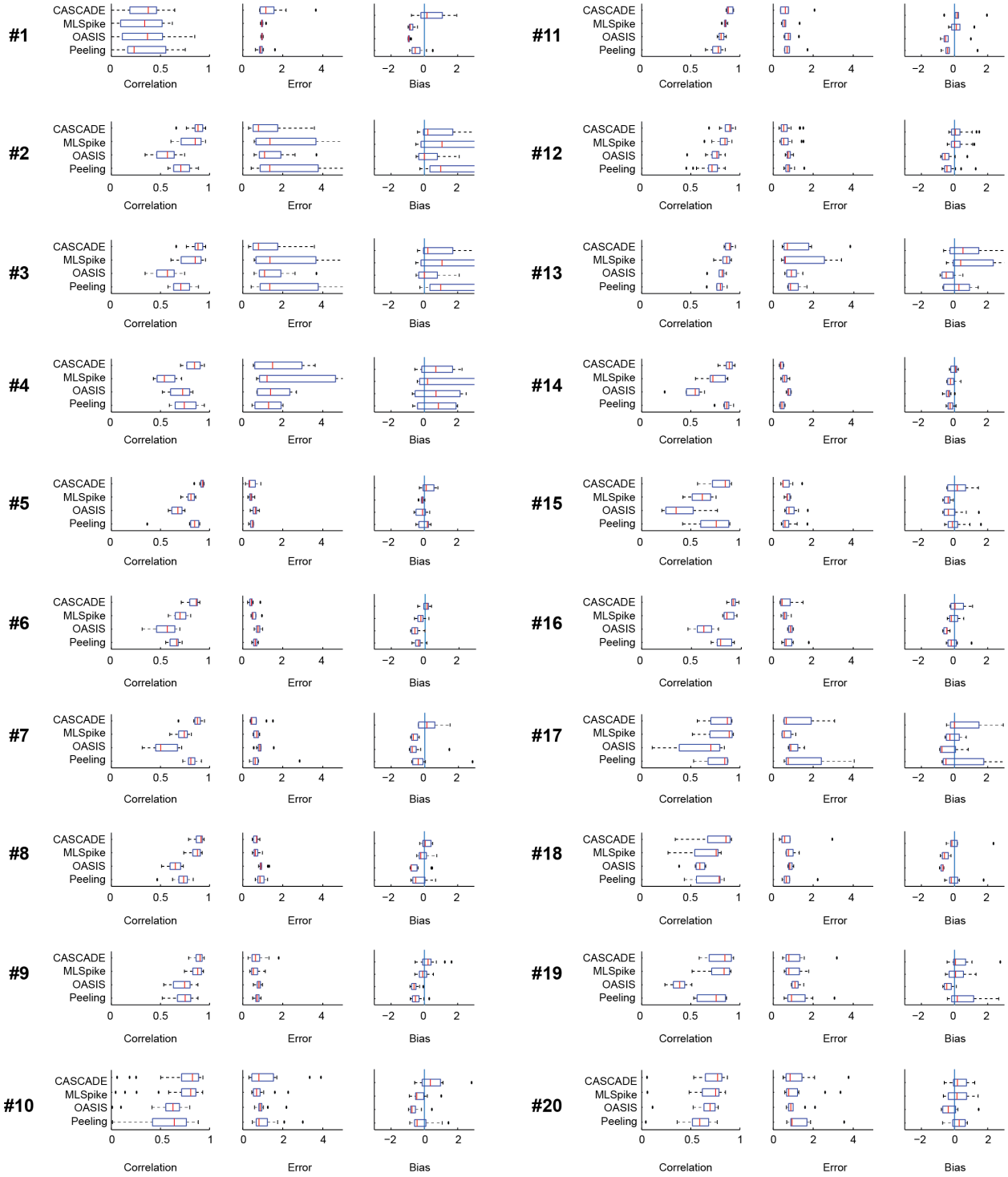

Figure S15 | Comparison of CASCADE with model-based algorithms, extension of Fig. 5b-e. Comparison of the four algorithms when optimized for a single dataset, with correlation (left), error (middle) and bias (right), for all datasets.

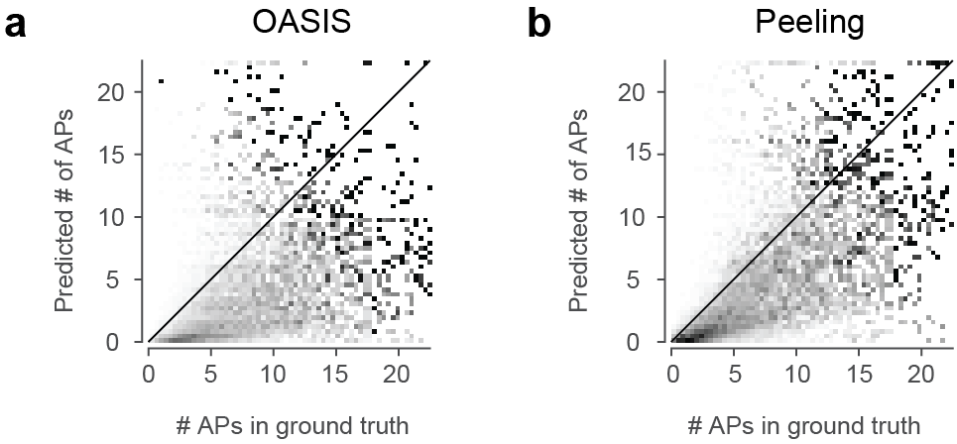

Figure S16 | **Bias of predictions across spike rates.** Extension of Fig. 5i for the model-based algorithms OASIS **(a)** and Peeling **(b)**.

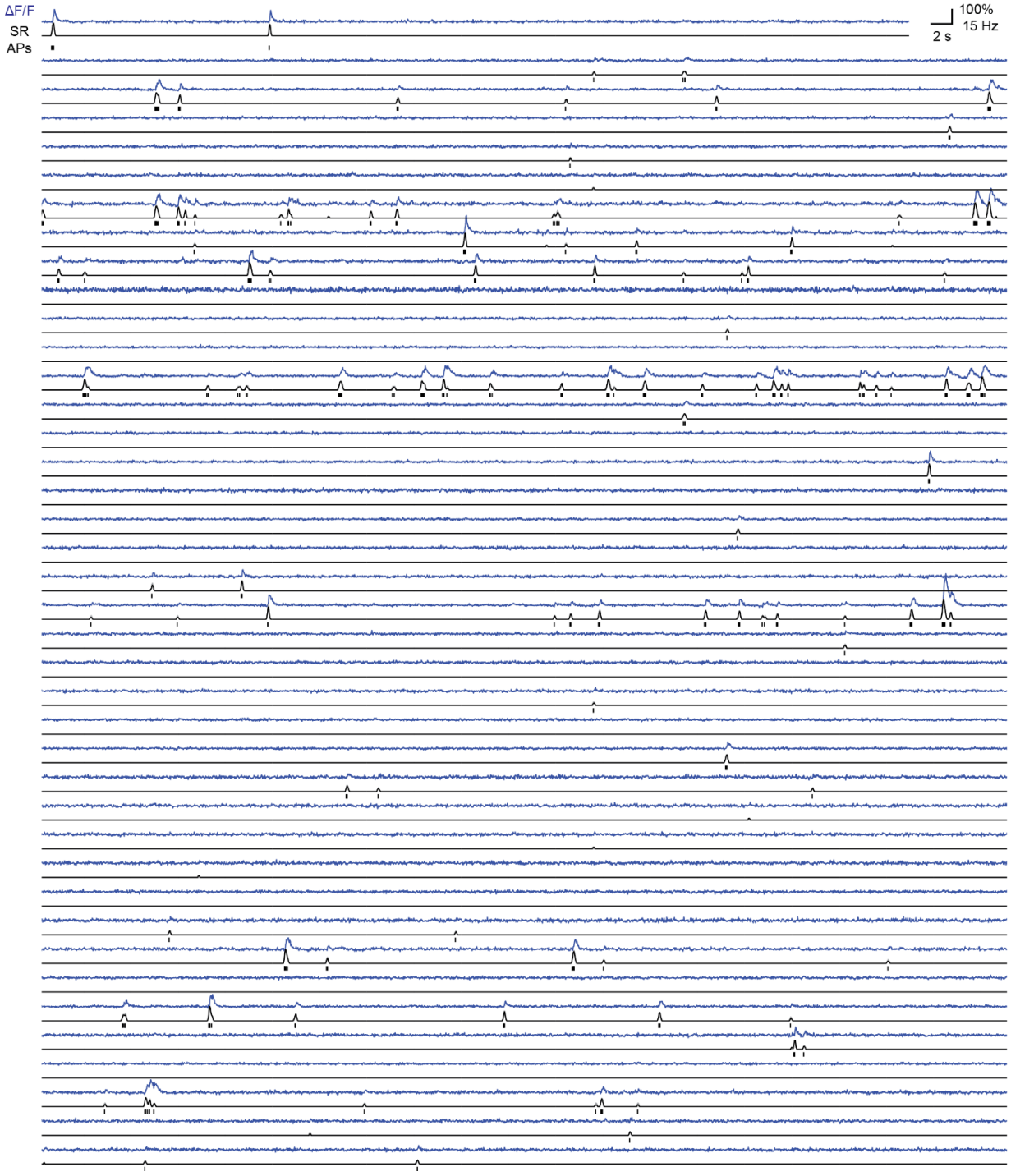

Figure S17 | Predictions of spiking probabilities and discrete spikes from the Allen Brain Institute Visual Coding dataset. From dataset ID '552195520', plotting a total of 40 neurons out of 74, approximately 1 minute out of 63.2 minutes of recording for this dataset. Discrete spikes are the most likely fit, generated with an algorithm using Metropolis-Monte Carlo sampling as starting point (see Methods).

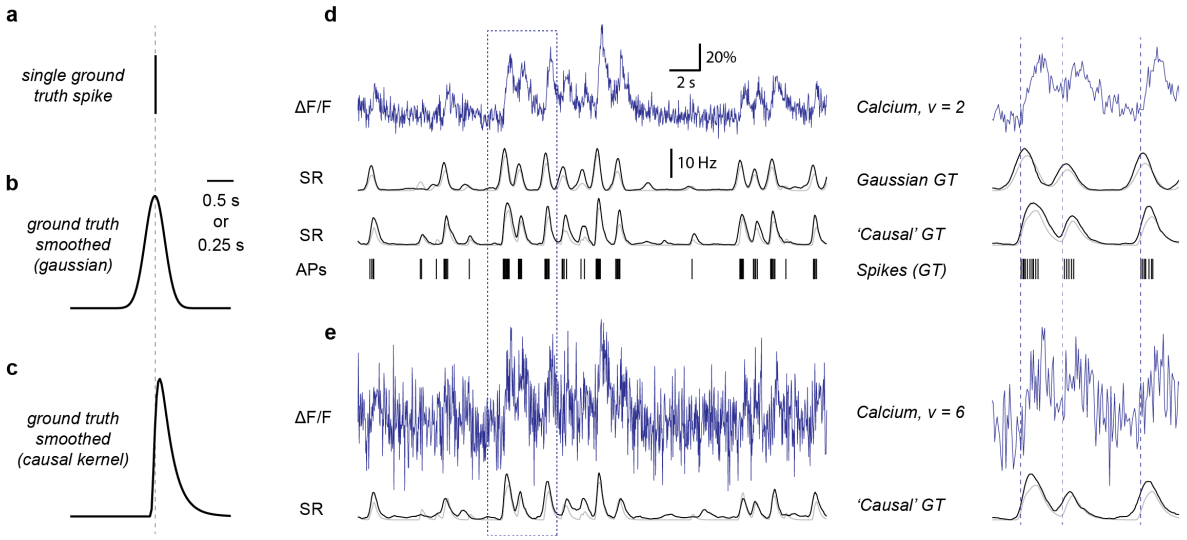

Figure S18 | **Non-Gaussian ground truth smoothing kernel.** Since calcium signals rise after action potentials, the task for spike inference is to re-assign ΔF/F activity to the true spike time. Due to noise and imperfect match between model and calcium trace, the re-assignment is typically slightly off, both into 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 this a-causal events, it is possible to use a more causal filter (**(c)**, 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 using a causal kernel to smooth the ground truth, a-causal effects can be induced by any spike inference algorithm. Depending on the desired temporal resolution of the algorithm, different width of the smoothing kernels can be chosen. A Gaussian with σ = 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 σ = 0.2 s was used.

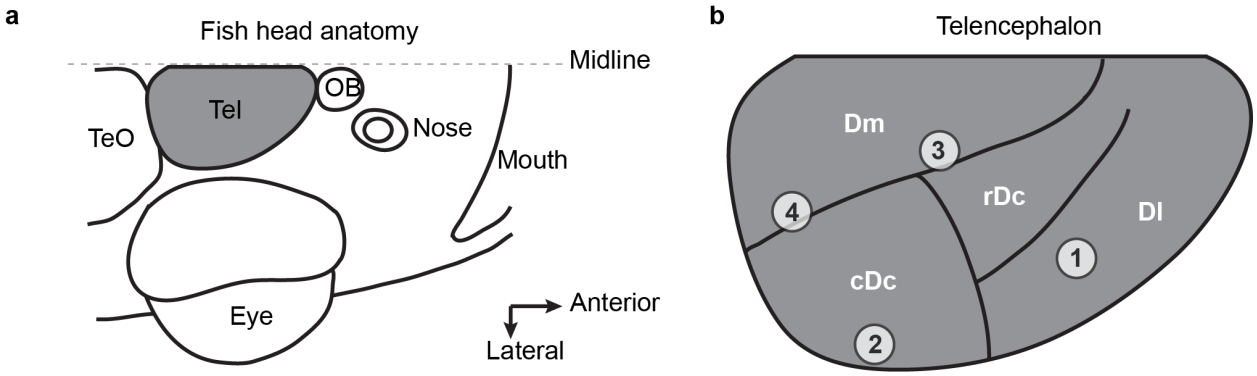

Figure S19 | **Locations of neurons in the dorsal telencephalon of adult zebrafish from ground truth dataset DS#06. a,** Fish head anatomy, highlighting the Telencephalon from a dorsal aspect. TeO = optic tectum, OB = olfactory bulb, Tel = Telencephalon. **b,** Recording locations in the telencephalon (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).

	Peeling		Peeling (corr)		MLSpike		MLSpike (corr)		OASIS	OASIS (corr)
Dataset	**τ (s)**	**A (a.u.)**	τ (s)	A (a.u.)	**τ (s)**	**A (a.u.)**	τ (s)	A (a.u.)	**τ (s)**	τ (s)
#1	0.75	5.0	0.5	2.5	0.75	0.35	1.5	0.01	0.02	0.82
#2	1.0	15	1.5	2.5	1.5	0.19	1.5	0.15	0.02	0.92
#3	1.0	15	1.5	2.5	1.5	0.19	1.5	0.15	0.02	0.92
#4	1.5	35	1.0	15	4.0	0.01	1.0	0.35	0.72	0.9
#5	2.5	35	2.5	10	2.0	0.09	2.5	0.09	0.82	0.96
#6	2.5	35	2.0	30	2.0	0.23	2.0	0.17	0.84	0.92
#7	2.0	7.5	2.5	10	1.0	0.03	2.5	0.35	0.18	0.94
#8	0.75	30	2.0	2.5	1.0	0.25	1.0	0.35	0.44	0.9
#9	1.0	35	0.5	30	0.75	0.17	0.75	0.17	0.56	0.8
#10	0.5	30	0.5	15	0.75	0.13	1.0	0.29	0.12	0.76
#11	2.5	35	0.75	10	3.0	0.35	1.5	0.35	0.78	0.86
#12	3.0	35	1.0	20	2.0	0.35	1.5	0.33	0.82	0.88
#13	5.0	35	2.5	25	5.0	0.35	2.5	0.35	0.96	0.94
#14	2.5	15	2.5	20	1.5	0.03	2.0	0.21	0.6	0.94
#15	2.0	10	2.5	10	0.75	0.03	2.0	0.35	0.18	0.94
#16	1.5	20	1.5	10	1.0	0.07	1.5	0.29	0.52	0.92
#17	0.5	15	0.5	2.5	0.75	0.07	1.0	0.29	0.02	0.86
#18	0.5	20	0.5	10	1.25	0.11	1.25	0.35	0.02	0.86
#19	2.0	10	2.0	5.0	2.5	0.11	2.5	0.35	0.12	0.94
#20	0.5	35	0.25	10	0.5	0.17	0.5	0.35	0.02	0.72

Table 2 (supplementary material) | 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). Black columns contain parameters when optimizing for the mean squared error of predictions. Grey columns contain parameters when optimizing for the correlation between predictions and ground truth.