



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

Peter Rupprecht<sup>1,2</sup>✉, Stefano Carta<sup>1</sup>, Adrian Hoffmann<sup>1</sup>, Mayumi Echizen<sup>3,4</sup>, Antonin Blot<sup>5,6</sup>, Alex C. Kwan<sup>7</sup>, Yang Dan<sup>8</sup>, Sonja B. Hofer<sup>5,6</sup>, Kazuo Kitamura<sup>3,9</sup>, Fritjof Helmchen<sup>1,11</sup>✉ and Rainer W. Friedrich<sup>2,10,11</sup>✉

**Inference of action potentials ('spikes') from neuronal calcium signals is complicated by the scarcity of simultaneous measurements of action potentials and calcium signals ('ground truth'). In this study, we compiled a large, diverse ground truth database from publicly available and newly performed recordings in zebrafish and mice covering a broad range of calcium indicators, cell types and signal-to-noise ratios, comprising a total of more than 35 recording hours from 298 neurons. We developed an algorithm for spike inference (termed 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; therefore, no parameters need to be adjusted by the user. In addition, we developed systematic performance assessments for unseen data, openly released a resource toolbox and provide a user-friendly cloud-based implementation.**

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<sup>1,2</sup>. However, calcium signals are only an indirect, often non-linear and low-pass filtered, proxy of the more fundamental variable of interest—that is, the train of somatic action potentials (spikes)<sup>3–6</sup>. The relationship between calcium signals and spike trains 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 inferring spike times or spike rates from other calcium imaging data (Fig. 1a). Based on such ground truth datasets, various model-based methods<sup>7–17</sup> as well as supervised machine learning algorithms<sup>16,18–21</sup> 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 neuron type, calcium indicator type and concentration, optical resolution, sampling rate and noise level. Many of these parameters can vary substantially between experiments and even among neurons 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

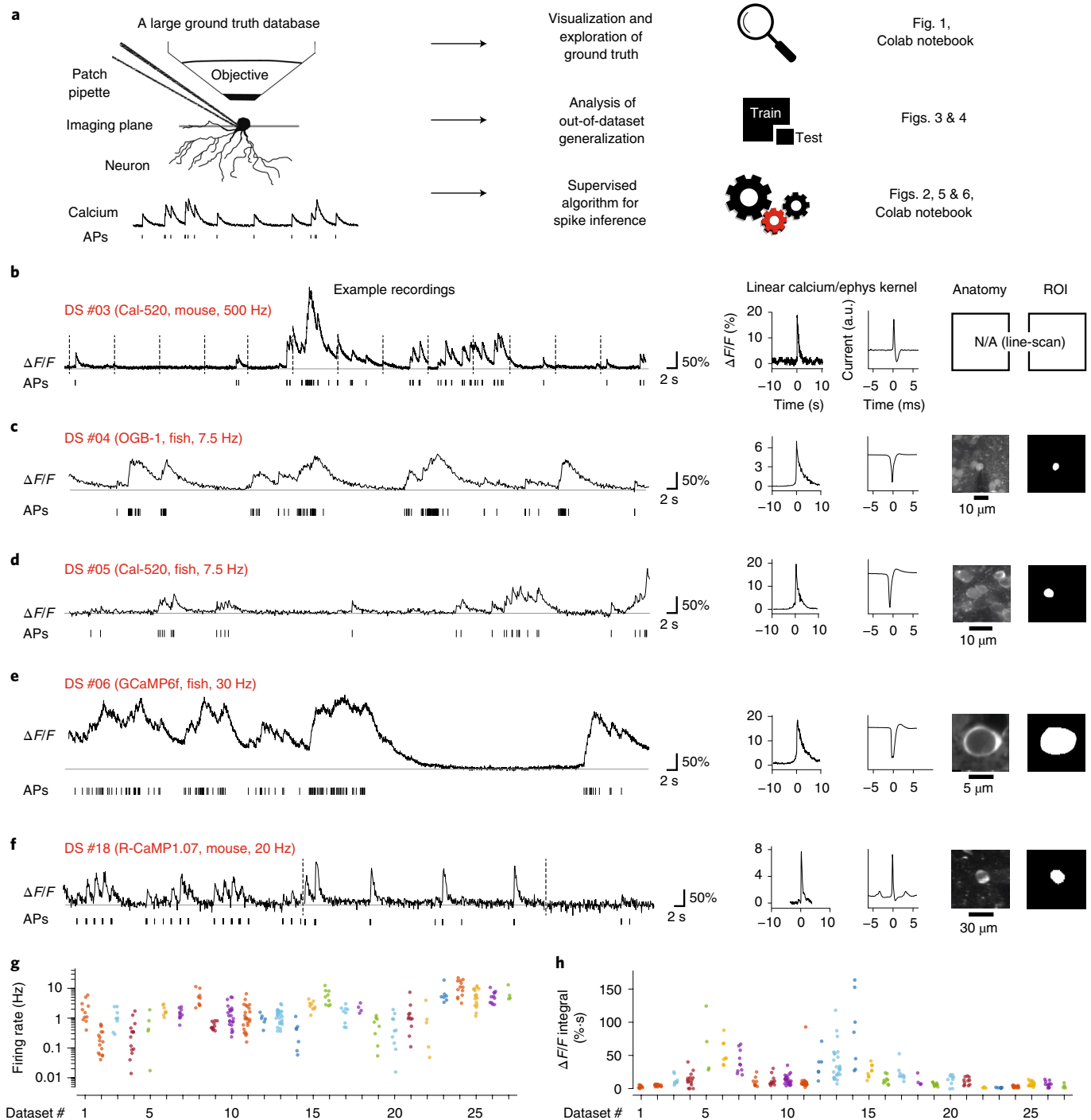
complicates the inference of spike rates from calcium imaging data under most experimental conditions<sup>13,14,22,23</sup>.

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 calibrated spike inference of calcium data using deep networks (termed CASCADE). CASCADE includes methods to resample the original ground truth datasets 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 outperformed existing algorithms in benchmark tests across all ground truth datasets and noise levels. 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

<sup>1</sup>Brain Research Institute, University of Zürich, Zurich, Switzerland. <sup>2</sup>Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. <sup>3</sup>Department of Neurophysiology, University of Tokyo, Tokyo, Japan. <sup>4</sup>Department of Anesthesiology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan. <sup>5</sup>Sainsbury Wellcome Centre for Neural Circuits and Behaviour, University College London, London, United Kingdom. <sup>6</sup>Biozentrum, University of Basel, Basel, Switzerland. <sup>7</sup>Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA. <sup>8</sup>Division of Neurobiology, Department of Molecular and Cell Biology, Helen Wills Neuroscience Institute, Howard Hughes Medical Institute, University of California, Berkeley, Berkeley CA, USA. <sup>9</sup>Department of Neurophysiology, University of Yamanashi, Yamanashi, Japan. <sup>10</sup>University of Basel, Basel, Switzerland. <sup>11</sup>These authors contributed equally: Fritjof Helmchen, Rainer W. Friedrich. ✉e-mail: [rupprecht@hifo.uzh.ch](mailto:rupprecht@hifo.uzh.ch); [helmchen@hifo.uzh.ch](mailto:helmchen@hifo.uzh.ch); [rainer.friedrich@fmi.ch](mailto:rainer.friedrich@fmi.ch)



**Fig. 1 | Ground truth datasets.** **a**, A large and diverse ground truth database obtained by simultaneous calcium imaging and juxtacellular recording (left) can be used (1) for the exploration of the ground truth by a user, (2) for the analysis of the out-of-dataset generalization of spike inference and (3) for the training of a supervised algorithm for spike inference. The right column refers to relevant figures. Collaboratory 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 ( $\Delta F/F$ ) are shown together with the detected action potentials. Dashed lines indicate breaks during recordings. Traces are representative for recordings from different datasets (see Table 1 for detailed information). Middle, linear kernels of  $\Delta F/F$  (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). Twenty-seven datasets were included in total. Datasets from inhibitory neurons comprise DSs #22–27. **h**, Integral  $\Delta F/F$  of the spike kernel (first 2 s) for each neuron. Lowest values are observed in PV-positive interneurons (DS #23 and DS #24). See Extended Data Fig. 1 for the underlying kernels. AP, action potential.

simultaneous electrophysiological recordings and calcium imaging in adult zebrafish and mice (Fig. 1b–h and 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<sup>24</sup> using the synthetic calcium indicators Oregon Green BAPTA-1 (OGB-1) and Cal-520 as well as the genetically encoded

**Table 1 | Overview of all ground truth datasets**

Dataset identifier	Calcium indicator	Induction method	Animal model	Brain region	Frame rate (Hz)	Standardized noise (%·Hz <sup>-1/2</sup> )	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	OGB-1	Injection + tg(tdTomato-CaMKII $\alpha$ )	Mouse	V1	15.6	0.5 ± 0.1	0.2 ± 0.2	16	116	Kwan and Dan, 2012
#3	Cal-520	Acute injection	Mouse	S1	500.0	0.3 ± 0.1	1.2 ± 0.8	8	23	Tada et al., 2014
#4	OGB-1	Acute injection	Zebrafish	pDp	7.7	1.0 ± 0.2	0.4 ± 0.5	15	81	This paper
#5	Cal-520	Acute injection	Zebrafish	pDp	7.8	2.0 ± 1.3	1.3 ± 2.1	5	31	This paper
#6	GCaMP6f	tg(NeuroD)	Zebrafish	aDp	30.0	1.3 ± 0.8	1.9 ± 0.7	8	46	This paper
#7	GCaMP6f	tg(NeuroD)	Zebrafish	dD	30.0	0.6 ± 0.1	1.5 ± 0.6	10	69	This paper
#8	GCaMP6f	tg(NeuroD)	Zebrafish	OB	30.0	0.8 ± 0.2	5.3 ± 3.3	9	45	This paper
#9	GCaMP6f	AAV	Mouse	V1	60.1	0.4 ± 0.1	0.6 ± 0.2	11	129	Chen et al., 2013
#10	GCaMP6f	tg(Emx1)	Mouse	V1	160.1	0.5 ± 0.2	1.6 ± 1.4	23	72	Huang et al., 2019
#11	GCaMP6f	tg(Cux2)	Mouse	V1	158.3	0.5 ± 0.2	1.5 ± 1.5	25	78	Huang et al., 2019
#12	GCaMP6s	tg(tetOs)	Mouse	V1	151.6	0.8 ± 0.1	1.0 ± 0.4	6	13	Huang et al., 2019
#13	GCaMP6s	tg(Emx1)	Mouse	V1	157.5	0.5 ± 0.2	1.3 ± 0.7	26	62	Huang et al., 2019
#14	GCaMP6s	AAV	Mouse	V1	60.1	0.5 ± 0.2	0.4 ± 0.4	7	70	Chen et al., 2013
#15	GCaMP6s	AAV	Mouse	V1	59.1	0.7 ± 0.2	6.2 ± 3.5	9	77	Theis et al., 2016
#16	GCaMP6s	AAV	Mouse	V1	59.1	0.9 ± 0.2	5.8 ± 3.3	9	25	Theis et al., 2016
#17	GCaMP5k	AAV	Mouse	V1	50.0	0.5 ± 0.2	1.6 ± 0.9	9	29	Akerboom et al., 2012
#18	R-CaMP1.07	tg(Grik4-cre) + AAV	Mouse	CA3	20.0	1.6 ± 0.3	2.2 ± 0.8	4	33	Schoenfeld et al., 2021
#19	R-CaMP1.07	AAV	Mouse	S1	15.0	0.6 ± 0.2	0.9 ± 1.0	9	50	Bethge et al., 2017
#20	jRCaMP1a	AAV	Mouse	V1	15.0	1.3 ± 0.5	0.6 ± 0.6	10	88	Dana et al., 2016
#21	jRGECO1a	AAV	Mouse	V1	29.8	1.0 ± 0.3	1.6 ± 2.0	11	118	Dana et al., 2016
#22	OGB-1	Injection + tg(GFP-GIN)	Mouse	V1 (SST)	15.6	0.6 ± 0.1	1.1 ± 1.6	5	49	Kwan and Dan, 2012
#23	OGB-1	Injection + tg(tdTomato-PV)	Mouse	V1 (PV)	15.6	0.6 ± 0.1	6.9 ± 5.4	7	17	Kwan and Dan, 2012
#24	GCaMP6f	tg(PV-cre) + AAV	Mouse (in vitro)	V1 (PV)	26.6	0.5 ± 0.2	11.6 ± 5.4	13	215	Khan et al., 2018
#25	GCaMP6f	tg(SOM-cre) + AAV	Mouse (in vitro)	V1 (SST)	26.6	0.6 ± 0.2	5.9 ± 3.5	17	375	Khan et al., 2018
#26	GCaMP6f	tg(VIP-cre) + AAV	Mouse (in vitro)	V1 (VIP)	26.6	2.1 ± 1.3	5.5 ± 2.5	11	252	Khan et al., 2018
#27	GCaMP6f	tg(tdTomato-PV) + AAV	Mouse	V1 (PV)	30.0	2.8 ± 1.1	7.0 ± 4.1	4	30	This paper

Standardized noise of calcium signals ( $\Delta F/F$ ) was determined as described in the Methods. Frame rate is 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 indicator GCaMP6f. In head-fixed mice, ground truth recordings were performed under anaesthesia in hippocampal area CA3 using the genetically encoded indicator R-CaMP1.07 (ref. <sup>25</sup>).

Furthermore, we extracted ground truth from published studies of neurons in mouse primary somatosensory cortex (S1), using Cal-520 and R-CaMP1.07, respectively (total of 21 neurons)<sup>26,27</sup>, and

of inhibitory neurons in mouse primary visual cortex (V1), using OGB-1 *in vivo* and GCaMP6f in slices, respectively (total of 69 neurons)<sup>28,29</sup>. A small new *in vivo* dataset for parvalbumin (PV)-positive neurons using GCaMP6f (four neurons) complemented this dataset. In addition, we surveyed openly accessible datasets and extracted ground truth from raw movies (when available) or pre-processed calcium imaging data<sup>16,19,30–34</sup>. Rigorous quality control (Methods) reduced the original number from a total of 193 available neurons to 157 neurons. Together with our own recordings, we assembled 27 datasets comprising a total of 298 neurons, eight calcium indicators and nine brain regions in two species, totaling ~38 h of recording and 495,077 spikes.

Recording durations, imaging frame rates and spike rates varied greatly across ground truth datasets (Table 1). Typical spike rates spanned more than an order of magnitude, ranging from 0.4 Hz to 11.6 Hz, and frame rates varied between 7.7 Hz and >160 Hz (Table 1 and Fig. 1g). Using regularized deconvolution, we computed the linear  $\Delta F/F$  kernel evoked by the average spike and found that the area under the kernel curve varied substantially across datasets, even for data from the same indicator, and was substantially smaller for datasets with inhibitory neurons, especially for PV interneurons (Fig. 1h). Interestingly, kernels showed large diversity even across neurons within the same dataset (Fig. 1h and Extended Data Fig. 1), which 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 >1,000 data points for each category in classification tasks)<sup>35</sup>. Second, the cost function can easily be modified to optimize the metric of interest—for example, 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 Supplementary Fig. 1 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  $\Delta F/F$ ) around a time point  $t$  to infer the spike rate at  $t$ . Compared to two-dimensional image classification and object labeling<sup>35,36</sup>, requirements on computational hardware are low because datasets are small and the inference task is only one-dimensional (time). For example, ImageNet<sup>37</sup>, a dataset used for visual object identification and detection in the deep learning field, is typically used at a resolution of  $256 \times 256 = 65,536$  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  $\Delta F/F$  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

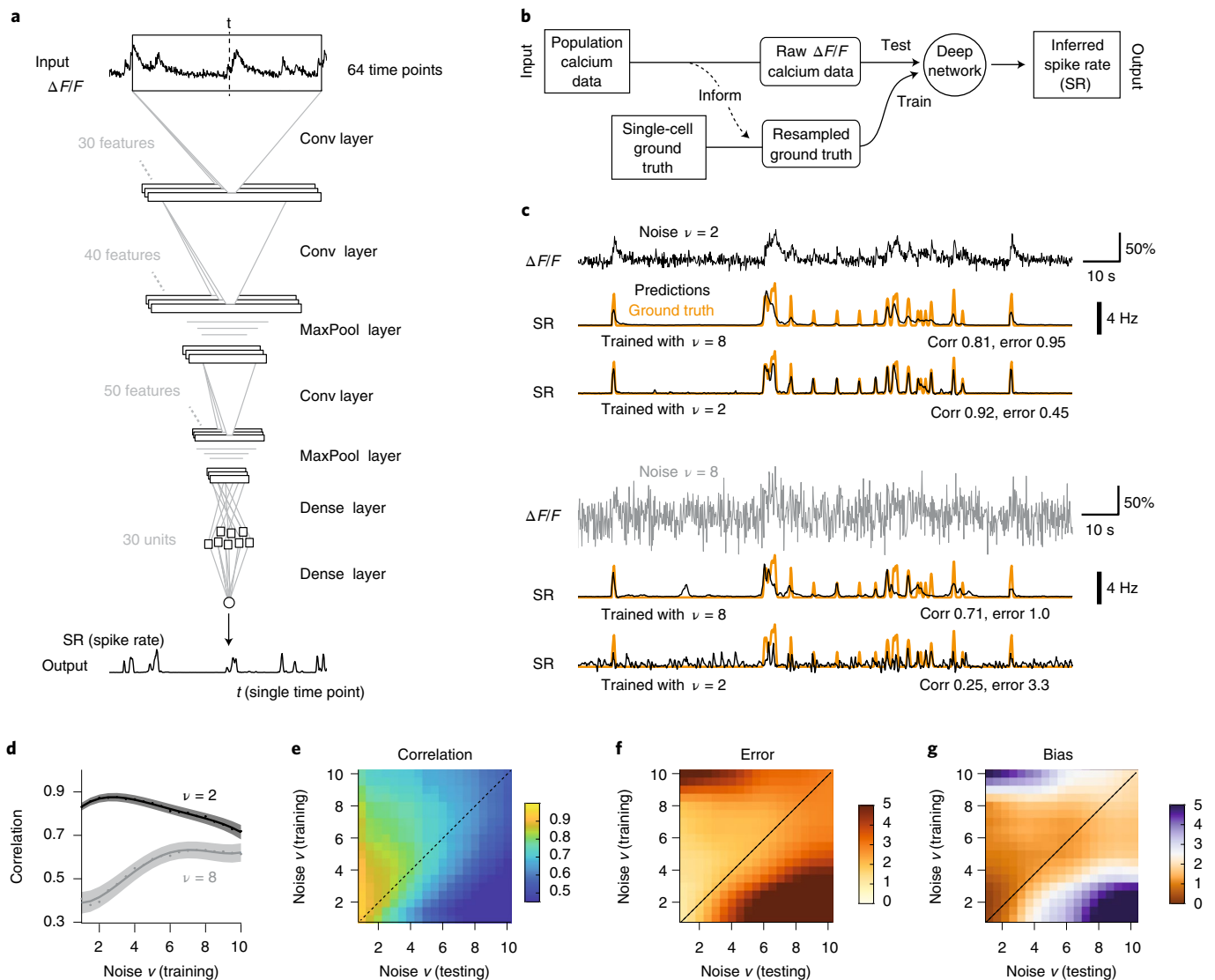
and then used to generate noise- and rate-matched training data from the ground truth database (Fig. 2b) by temporal resampling and addition of noise. To facilitate gradient descent, the ground truth spike rate is smoothed with a Gaussian kernel ( $\sigma = 0.2$  s unless otherwise indicated; Methods).

To extract  $\Delta F/F$  noise levels, we computed a standardized noise metric  $\nu$  that is robust against outliers and approximates the standard deviation of  $\Delta F/F$  baseline fluctuations. This metric was normalized by the square root of the frame rate to allow for comparison of noise measurements across datasets. Consequently,  $\nu$  has units of  $\% \cdot \text{Hz}^{-1/2}$ , which, for simplicity, we usually omit (Methods and Extended Data Fig. 2). To generate training data with pre-defined  $\Delta F/F$  noise levels, we explored several approaches based on sub-sampling of regions of interest (ROIs) or additive artificial noise (Supplementary Note 1 and Supplementary Fig. 2). 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 (equations and illustrations in Supplementary Fig. 3). Following previous studies, we calculated the Pearson correlation between ground truth spike rates and inferred spike rates<sup>16,19</sup>. As this correlation measure of performance leaves the absolute magnitude of the inferred spike rate unconstrained, we also determined two additional quantities: the error, which was defined as sum of the absolute deviations between the inferred spike rate and the ground truth, and the bias, which was defined as the sum of the signed deviations (Methods and Supplementary Fig. 3). Error and bias were both normalized by the number of true spikes to obtain relative metrics that can be compared between datasets. Of these three metrics (correlation, error and bias), correlation is arguably the most important one because it estimates the similarity of inferred and true spike rates. Error and bias are relevant for the inference of absolute spike rates because they identify spike rate estimates that are incorrectly scaled or systematically too large or small.

The performance of the deep network degraded considerably when the noise level of the test dataset deviated substantially from the noise level of the ground truth. As expected intuitively, a network that had only seen almost noise-free data during training failed to suppress fluctuations in noisier recordings. Conversely, we 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 and Extended Data Fig. 3). Very low noise levels ( $\nu < 2$ ) result in a special case (Fig. 2d,e): because some neurons of a given ground truth dataset do not reach the desired noise level even without addition of noise (compare to Table 1), the effective size of the training dataset decreases, resulting in slightly lower performance. In general, however, it turned out beneficial to train with noise levels that are adapted to the calcium data to which the algorithm will be applied after training.

**Parameter robustness of spike inference.** Traditional models to infer spiking activity typically contain a small number of parameters<sup>11–13,15</sup> 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<sup>14,16</sup>. 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—that is, the size of the input window shown in Fig. 2a. We, therefore, tested how spike inference performance depends on these hyper-parameters.



**Fig. 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 and 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  $\Delta F/F$  trace is translated into spike rates (inferred spike rates in black and ground truth in orange) more precisely when low-noise ground truth has been used for training. Bottom, a high-noise  $\Delta F/F$  trace is translated into spike rates (inferred spike rates in black and ground truth in orange) more precisely when high-noise ground truth has been used for training.  $\nu$  in units of standardized noise,  $\% \cdot \text{Hz}^{-1/2}$ . **d**, The spike inference performance for two test conditions (low noise,  $\nu = 2$ , dark gray; high noise,  $\nu = 8$ , light gray) 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. **f**, Relative error of predictions with respect to ground truth. **g**, Relative bias of predictions with respect to ground truth. Column-wise normalized versions of **e-g** are shown in Extended Data Fig. 3.

We found that the performance of the network was robust against variations of all hyper-parameters (Supplementary Note 2 and Supplementary Fig. 4a–e), allowing us to leave all parameters unchanged for all conditions. 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 and Supplementary Fig. 4f–h). Finally, we tested different deep learning architectures, including non-convolutional or recurrent long short-term memory (LSTM) networks. Although very large networks tended to slightly overfit the data, most networks performed almost equally well (Supplementary Note 2 and Supplementary Fig. 5). 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.

**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, labelling method, concentration levels, brain region and cell type). 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  $\Delta F/F$  data were resampled to a common sampling rate and adjusted to the same noise levels by adding Poisson noise. If the initial noise level of a given ground truth neuron was higher than the target noise level, the neuron was excluded from this analysis. We then evaluated the performance of CASCADE as a function of the noise levels of the (resampled) datasets. As expected, correlations increased and errors decreased for lower noise levels, whereas average biases were not systematically affected (Extended Data Fig. 4a–d). Performance metrics also varied considerably across different neurons within a single dataset when resampled at the same noise level  $\nu$ . 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. 1h and Extended Data Fig. 1). 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 Extended Data Fig. 4; compare to Extended Data Fig. 1r for the respective linear kernels of dataset (DS) #18).

Second, spike inference might be complicated by movement artifacts or neuropil contamination. Movement artifacts typically had slow onset and slow offset kinetics (Extended Data Fig. 5a) or a faster, quasi-periodic temporal structure related to breathing (Extended Data Fig. 5d,e). Neuropil contamination is often difficult to distinguish from somatic calcium signals and particularly severe when neurons are tightly packed and densely labelled<sup>1,38,39</sup> (Extended Data Fig. 5b). For a subset of datasets, we tested the effect of simple center-surround subtraction of the neuropil signal<sup>30</sup>. Because subtraction is not perfect, decontaminated datasets still contained residual neuropil signals (Extended Data Fig. 5b) or negative transients (Extended Data Fig. 5c). Nonetheless, spike inference was significantly improved by neuropil decontamination (Supplementary Fig. 6). More detailed inspection of the results showed that CASCADE learned to ignore negative transients and movement artifacts but only as long as they were distinguishable from true calcium transients (Extended Data Fig. 5a–c).

Third, we found that the activity of sparsely spiking neurons is less well predicted because the calcium signal of single action potentials is more likely to be overwhelmed by shot noise, particularly in the high-noise regime (arrows in Extended Data Fig. 4a,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 (Supplementary Fig. 7). This observation was made using GCaMP indicators, which show a strongly non-linear 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 (Supplementary Fig. 7). These observations indicate that the network needs to learn a tradeoff between false-positive detections of noise events and false-negative detections of single spikes. Additional details related to single-spike precision and the possibility to discretize inferred spike rates are discussed in Supplementary Note 3.

In summary, we found that CASCADE is able to generalize to unseen neurons from the same ground truth training set. Not surprisingly, the accuracy of generalization decreases with increasing noise levels, in particular when spike rates are low. Accuracy is fundamentally limited by the variability of calcium kernels across neurons and probably also by the non-linearity of GCaMP-like indicators, and accuracy is further reduced when additional noise (motion artifacts or neuropil contamination) is prominent.

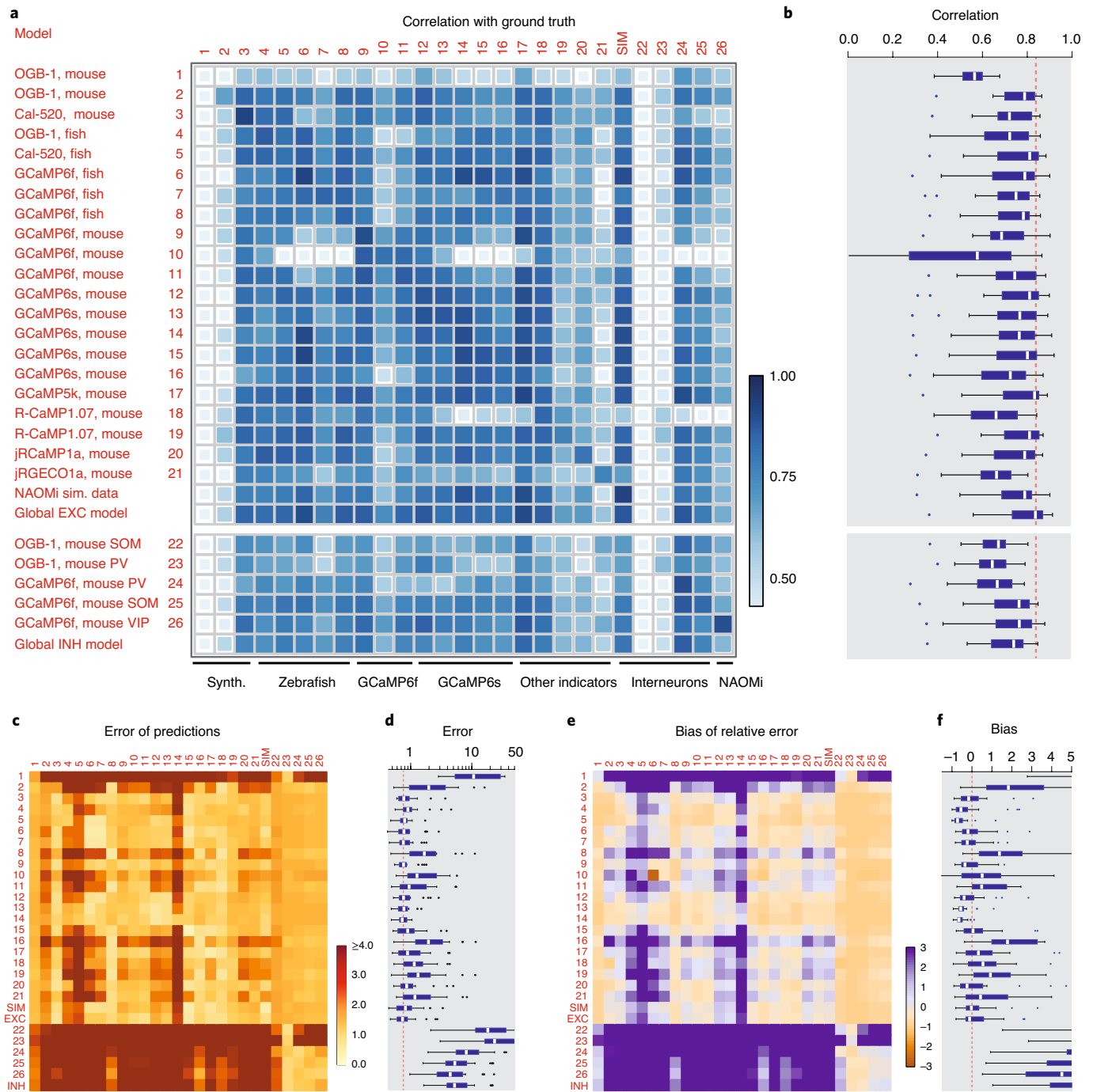
**Generalization across datasets.** We next explored how spike inference by a network trained on one ground truth dataset generalizes to other datasets. Using all available datasets, we quantified the median performance metrics across all possible combinations of datasets for training and testing and analyzed the performance of each trained model across test datasets (Fig. 3). In most training/test combinations, correlations were high, whereas errors and biases remained low. Exceptions were rare and occurred in datasets with considerable motion or neuropil contamination artifacts (for example, DS #01 and DS #02 and DSs #21–23). The entries of the matrix in Fig. 3 remained highly similar when parameters such as the resampling rate, temporal smoothing of the ground truth or the noise level were modified (Supplementary Fig. 8). Interestingly, models trained on datasets that were dominated by excitatory neurons (DSs #01–21, hence called ‘excitatory datasets’) also produced high-quality predictions of spike rate variations in inhibitory neurons (DSs #22–26, ‘inhibitory datasets’; Fig. 3a,b), although the separate analysis of error and bias revealed that absolute spike rates were substantially underestimated (Fig. 3c–f).

Near-maximal correlation for a given dataset was often achieved by multiple models (Fig. 3a). In some datasets, the highest correlation was even achieved when the model was trained on ground truth from another dataset. Interestingly, the performance of training/testing combinations showed no obvious clustering related to indicator type (for example, genetically encoded versus organic indicators) or species (zebrafish versus mouse). An attempt to explain the mutual predictability of datasets by more refined statistical dataset descriptors, such as the mean spike rate or decay times, was not very successful (Supplementary Fig. 9). 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 tested an alternative and simpler approach by training a model on all excitatory datasets except DS #01, hence called the ‘global EXC model’ (abbreviated as ‘EXC model’). We found that this global model performed better than all other models in cross-dataset tests (Fig. 3a–f; the test dataset was always excluded from training data), due not only to the size but also to the diversity of the training set (Extended Data Fig. 6). Compared to randomly selecting a single dataset with excitatory neurons for training, correlations were increased by  $0.05 \pm 0.05$ ; errors were reduced by  $0.05 \pm 0.05$ ; and absolute biases were reduced by  $0.25 \pm 0.90$  (median  $\pm$  s.d.). In addition, the global EXC model performed better than any of the 21 single models in all cross-dataset tests ( $P < 0.001$  for all comparisons, paired signed-rank test). Compared to predictions across neurons within the same dataset (Extended Data Fig. 4; diagonal elements in Fig. 3), the correlations resulting from the EXC model were decreased by  $0.02 \pm 0.04$  ( $P = 0.04$ , Wilcoxon signed-rank test), and errors were increased by  $0.33 \pm 0.53$  ( $P = 0.01$ ), whereas the absolute bias was slightly decreased ( $0.40 \pm 0.40$ ,  $P = 0.002$ ). Hence, using dataset-specific ground truth can yield performance significantly better than the global EXC model. In the absence of such specific calibration data, however, training the algorithm with all available data is a simple and effective strategy to generate a model that generalizes robustly to unseen datasets.

Not surprisingly, a global ‘INH model’ trained on all inhibitory datasets (DSs #22–26) generalized less well across all datasets than the EXC model (Fig. 3). Indeed, the INH model was not more successful than the EXC model in predicting activity of inhibitory neurons with respect to correlation or error ( $P = 0.84$  and  $P = 0.68$ ; Fig. 3a,c), although the bias was lower ( $P = 0.03$ ; Fig. 3e). Most likely, generalization to unseen inhibitory neurons could be further improved by additional ground truth for inhibitory neurons.

We also trained a model on a large artificial dataset (250 neurons) that was generated using the calcium imaging simulation environment NAOMi<sup>40</sup> (Methods). The model performed well but lower



**Fig. 3 | 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. ‘NAOMi’ is a model trained on simulated GCaMP6f data based on Charles et al.<sup>40</sup>. Rows 21–24 are networks trained on datasets with inhibitory neurons. ‘Global EXC model’ and ‘global INH model’ are globally trained on all excitatory or inhibitory datasets (except DS #01 and the respective test dataset). **a**, Correlation of predictions with the ground truth. The size and color of the squares scale with correlation. **b**, Distribution of the performance of each trained network (row) across all other datasets (distribution across  $n=25$  datasets for each box plot). The dashed line highlights the median of the best-performing model (‘global EXC model’). **c**, **d**, Relative error of predictions compared to the ground truth. The dashed line in **d** highlights the median of the best-performing model (‘global EXC model’). **e**, **f**, Relative bias of predictions compared to the ground truth (distribution across  $n=25$  datasets for each box plot). All datasets were resampled at a frame rate of 7.5 Hz, with a standardized noise level of 2. For box plots, the median is indicated by the central line; 25th and 75th percentiles are indicated by the box; and maximum/minimum values excluding outliers (points) are indicated by the whiskers.

than the global EXC model (correlation reduced by  $0.05 \pm 0.04$ ,  $P=0.0003$ ; error slightly increased by  $0.06 \pm 0.22$ ,  $P=0.0006$ ; bias not significantly changed,  $P=0.67$ ; Fig. 3). We hypothesize that some relevant sources of variability at the neuronal level (for

example, variable decay times, transient shapes and non-linearities) are captured by experimental ground truth but not by simulated ground truth recordings. A future application of NAOMi could be the simulation of ground truth data for new calcium indicators,

when biophysical parameters are known but experimental ground truth is not available.

**Comparison with existing methods.** To benchmark the performance of CASCADE, we compared it to five other model-based methods: the fast online deconvolution procedure OASIS with two distinct implementations in CaImAn and Suite2p<sup>15,38,41</sup>; the discrete change point detection algorithm by Jewell and Witten<sup>42</sup> (here referred to as Jewell&Witten); and two more complex algorithms, Peeling and MLSpikes. Peeling uses iterative template subtraction to infer discrete spikes<sup>11</sup>. MLSpikes was chosen because it outperformed various other methods in previous applications<sup>12,16</sup>. Although model-based methods are, in principle, non-supervised, several parameters need to be tuned to achieve maximal performance on a given dataset<sup>13</sup>. To avoid sub-optimally tuned algorithms and to make the comparison with CASCADE as fair as possible, we used extensive grid searches to optimize parameter tuning of each algorithm/dataset combination (Methods; see Supplementary Table 1 for the best model parameters for each dataset as a function of noise). This procedure allowed us to minimize the same loss function for all algorithms (mean squared error between ground truth and the inferred spike rate), using grid search for model-based approaches and back-propagation for CASCADE. Notably, the neuron used for testing was always omitted during the training/fitting period (leave-one-out strategy). We refer to these models as 'tuned' for specific datasets, as opposed to CASCADE's 'global EXC model' that was trained on other datasets (Fig. 3). The Peeling and Jewell&Witten algorithms infer discrete spikes rather than spike rates, which might result in a slight disadvantage. To convert their output to continuous rates, predicted spikes were convolved with a Gaussian kernel of a width that minimized the mean squared error.

The tested algorithms showed systematic differences in performance (Fig. 4a and Extended Data Fig. 7). A quantitative comparison across all datasets for a fixed noise level revealed that performance varied strongly across ground truth datasets, single neurons and algorithms (Fig. 4b and Extended Data Fig. 8). Neurons that could be predicted well by one algorithm could often also be predicted well by other algorithms (see Extended Data Fig. 8 for error and bias), suggesting that outlier neurons within datasets exhibit unusual properties that lead to biased predictions (Extended Data Fig. 4). The tuned CASCADE model and CASCADE's EXC model produced good predictions for the broadest set of neurons across datasets. High-level performance of the model-based algorithms was observed in fewer datasets. For example, in multiple neurons from diverse datasets, the performance of MLSpikes was lower than CASCADE (Fig. 4b; DS #7 and DS #8 (GCaMP6f in fish), DS #15 and DS #16 (GCaMP6s in V1) and DSs #24–26 (GCaMP6f in inhibitory neurons)). These datasets had relatively high (Table 1) and slowly changing spike rates rather than discrete bursts. Interestingly, the Peeling algorithm performed relatively well on some of these datasets. To more directly compare the performances across neurons with CASCADE, we calculated the difference in correlation achieved by CASCADE and other algorithms for each neuron. The resulting distributions (Fig. 4c) show that CASCADE yielded better inferences for most neurons across all compared algorithms ( $P < 10^{-10}$  for all comparisons with other algorithms;  $P = 0.068$  when compared to CASCADE's global EXC model; paired Wilcoxon signed-rank test).

The performance of CASCADE was consistently better across different recording conditions. First, based on the finding that noise levels affect spike inference more strongly than other parameters (Supplementary Fig. 8), we repeated the benchmarking in Fig. 4b across multiple noise levels. Performance ranking across algorithms was largely maintained (Fig. 4d), with the global EXC model achieving performance close to the tuned CASCADE model (significant difference:  $P = 0.039$ , signed-rank test), followed by MLSpikes, Peeling, Suite2p and CaImAn and then followed by Jewell&Witten

( $P < 10^{-10}$  for all algorithms). Although the error computed from CASCADE's predictions was significantly lower than for most other algorithms ( $P < 0.005$  for Jewell&Witten, Suite2p, MLSpikes and CASCADE's EXC model;  $P < 0.01$  for CaImAn but  $P > 0.05$  for Peeling; paired Wilcoxon signed-rank test), variability was high and relative effect sizes were low (Fig. 4e). Therefore, errors are not a very sensitive readout of performance. Finally, biases of predictions were negative (indicating underestimates of true spike rates) for all tuned model-based algorithms except for the tuned CASCADE model (Fig. 4f). CASCADE's EXC model exhibited the smallest overall bias.

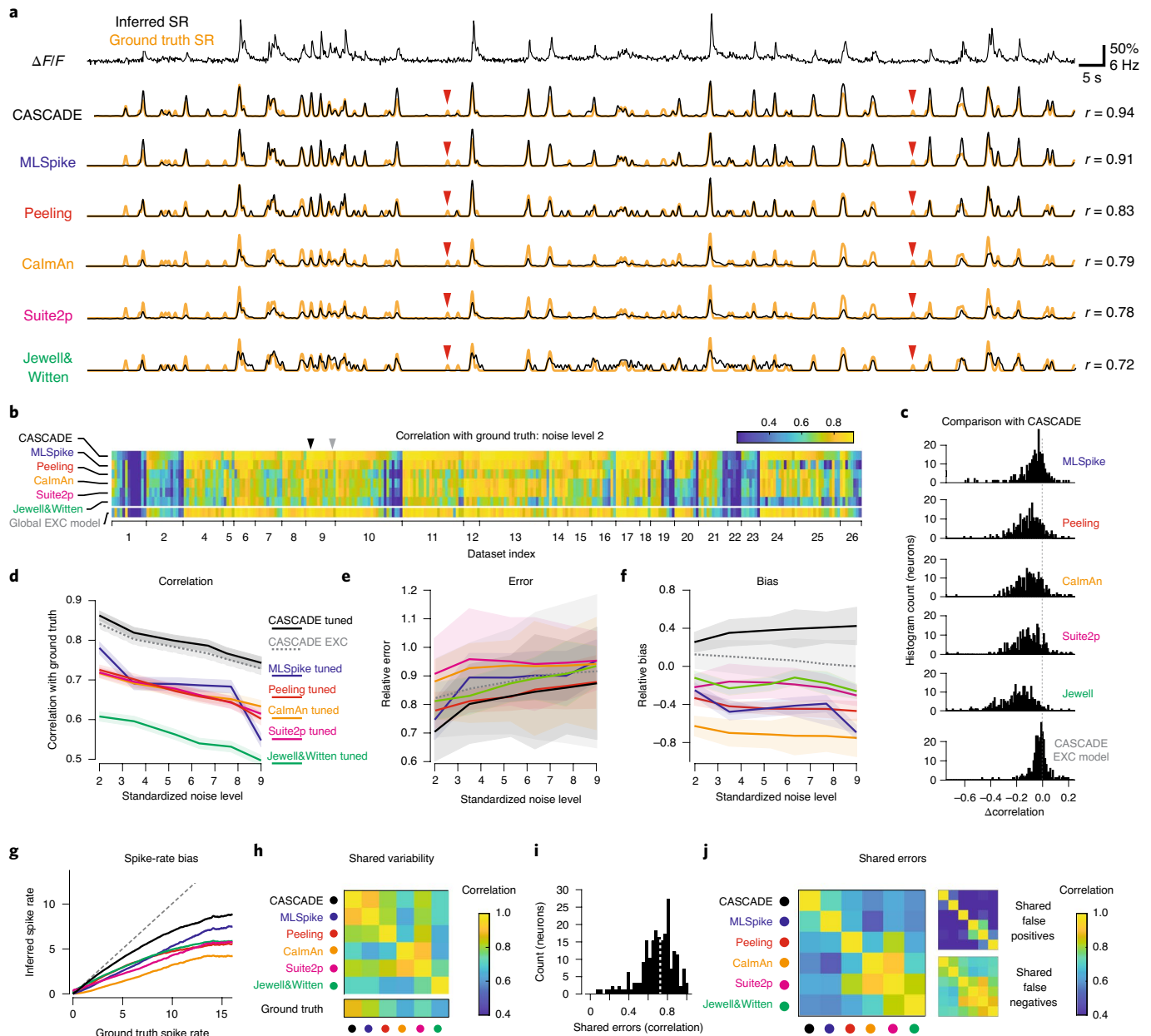
We further found that all algorithms systematically underestimated high spike rates. This effect was, on average, smallest for CASCADE. To visualize these results, we plotted the number of spikes for ground truth and predictions within each 2-s time bin (Supplementary Fig. 10) and extracted the median lines of these distributions (Fig. 4g). An underestimate of high spike rates might be expected because periods of high activity are rare; false-positive predictions of high spike rates might, thus, lead to larger performance drops than false-negative omissions of rare events.

For spike inference evaluated at higher temporal precision, the performance (correlation with ground truth) dropped for all algorithms, but this effect was more modest for CASCADE than for all model-based algorithms (Extended Data Fig. 9). We trained all algorithms to a ground truth that was smoothed in time to a variable degree (Gaussian smoothing kernel between  $\sigma = 0$  ms and  $\sigma = 333$  ms; default:  $\sigma = 200$  ms). Example predictions highlight that several algorithms make impressive predictions also under these more difficult conditions (Extended Data Fig. 9a), but some algorithms, in particular those based on discrete events (Peeling and Jewell&Witten), were not able to include graded certainties about spike times and, therefore, performed less well (Extended Data Fig. 9b). However, also the performance of MLSpikes, CaImAn and Suite2p dropped faster than the performance of CASCADE when spike rates were evaluated with increasing temporal precision (Extended Data Fig. 9c).

Predictions of different algorithms were not only similar across neurons (Fig. 4b, Extended Data Fig. 8) but also exhibited correlated temporal structure for predictions from the same neurons (Fig. 4h–j). The shared variability, measured as the median correlation between predictions, was particularly high for the two closely related algorithms—Suite2p and CaImAn—but also for CASCADE and MLSpikes. Indeed, the correlation between CASCADE and MLSpikes was as high as the correlation between CASCADE and the ground truth (Fig. 4h, bottom). To better understand these similarities, we explored false predictions shared by algorithms and computed the similarities (correlation) of the unexplained, residual variances across algorithms. These shared errors were prominent (Fig. 4i,j). In particular, errors made by CaImAn, Suite2p, Peeling and Jewell&Witten were often correlated, but CASCADE and MLSpikes also shared a relatively large fraction of unexplained variance. We further divided the unexplained variance into false positives (predictions higher than the ground truth) and false negatives (predictions lower than the ground truth). False negatives but not false positives were highly correlated across most algorithms, with the exception of Suite2p and CaImAn, which also shared false positives (Fig. 4j, right). Shared false negatives are clearly visible in typical predictions (red arrows in Fig. 4a and, more prominently, in Extended Data Fig. 7). Together, these analyses show highly similar predictions and similar missed spike events across algorithms. In summary, CASCADE predicted spike rates more accurately than all other algorithms across datasets, across noise levels and for different temporal precisions. Moreover, CASCADE showed a smaller bias toward underestimating high spike rates.

Finally, we compared practical aspects arising during the application of different algorithms. With respect to processing speed,





**Fig. 4 | Comparison with model-based algorithms.** **a**, Example calcium imaging recording and corresponding predictions from the deep learning-based method (CASCADE) and five model-based algorithms (MLSpike, Peeling, CalmAn, Suite2p and Jewell&Witten). Respective predictions are in black, and ground truth is in orange. *r* indicates correlation of predictions with ground truth. Clear false-negative detections are labelled with red arrowheads. **b**, Heat map of the performance (correlation) of each algorithm for each dataset and neuron, calculated at standardized noise level 2% Hz<sup>-1/2</sup>. All algorithms, except for CASCADE's global EXC model (compare to Fig. 3), were tuned to the respective dataset by the mean squared error between ground truth and inferred spike rate. Arrowheads highlight the example neurons shown in Fig. 4a (black) and Extended Data Fig. 7 (gray). **c**, Direct comparison of performance (**b**) between CASCADE and other algorithms on a single-neuron basis. The difference in performance (correlation) is shown as a histogram across all neurons. 'Global EXC model' as defined in Fig. 3. **d-f**, Comparison of correlation, error and bias across all algorithms and noise levels. *ν* in units of standardized noise, %Hz<sup>-1/2</sup>. Solid/dashed lines indicate the mean across all neurons, and shaded areas represent the s.e.m. **g**, Spiking activity in 2-s bins, ground truth versus predictions. Lines indicate medians across distributions. Algorithms are color-coded as before. Underlying distributions are shown in Supplementary Fig. 10. The unity relationship is shown as a dashed line. **h**, Variability shared across algorithms, measured by the correlation between predictions. **i**, Histogram of error shared between CASCADE and MLSpike, quantified as the correlation between the unexplained variances for each neuron. The dashed line indicates the median. **j**, Shared median errors as illustrated in **i** for all pairs of algorithms. The smaller matrices to the right break the shared errors down into false positives and false negatives. All quantifications were performed with ground truth datasets resampled at 7.5Hz with a noise level of 2 unless otherwise indicated. DS #03 was omitted for all comparisons in Fig. 4 because the short recordings (<10s) could not be processed by all algorithms. SR, spike rate.

we found that CASCADE (based on a GPU), CalmAn and Jewell&Witten performed similarly fast (200,000–300,000 samples per second). They were outperformed only by Suite2p (more than

5 million samples per second), whereas Peeling (5,000 samples per second) and, in particular, MLSpike (800 samples per second) were much slower. For optimization, CASCADE uses back-propagation,

which is almost equally fast as inference, resulting in a total training time of less than 10 min for a typical ground truth dataset with 5 million data points and a realistic number of 20 iterations (epochs) through the dataset. For model-based algorithms, we performed extensive grid searches across parameters (usually in a two-dimensional parameter space with 100–500 parameter combinations), which is feasible within minutes for Suite2p, CalmAn and Jewell&Witten. For Peeling and MLSpoke, this procedure would take several days for a single model. We, therefore, reduced the number of training samples for MLSpoke and Peeling to achieve search times of approximately 2 h per model for MLSpoke. Furthermore, we found that best fit parameters for model-based approaches changed systematically with noise levels, suggesting that new models have to be fit for each noise level (Supplementary Table 1), an effect that was more pronounced for algorithms that do not use the noise level as an input (Suite2p and Jewell&Witten). For some model-based algorithms, we found that inferred spike rates were often temporally shifted to later time points, and this delay was variable across datasets ( $0.16 \pm 0.14$  s for MLSpoke,  $0.03 \pm 0.09$  s for Peeling,  $0.31 \pm 0.23$  s for CalmAn,  $0.29 \pm 0.22$  s for Suite2p and  $0.27 \pm 0.19$  s for Jewell&Witten; mean delay  $\pm$  s.d. across datasets). We corrected these shifts for all analyses presented here. Such a correction is not necessary for a supervised algorithm such as CASCADE, which learns the correct shift from the ground truth. Together, these aspects reflect that, unlike model-based algorithms, CASCADE can make use of ground truth datasets in an efficient and natural way.

**Application to population calcium imaging datasets.** A transformation of calcium signals into estimates of spike rates might be desired for multiple reasons. First, the reconstruction of spike rates can recover fast temporal structure in neuronal activity that is obscured by slower calcium signals<sup>4,5</sup>. Second, a method that infers spiking but ignores noise can eliminate shot noise and potentially other forms of noise without the detrimental effects of over-expressed indicators<sup>1</sup> and without compromising temporal resolution. Third, although 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<sup>24</sup>, 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<sup>43</sup> was performed as in DS #04 at a noise level of  $2.36 \pm 0.97$  (%-Hz<sup>-1/2</sup>; median  $\pm$  s.d.) across 1,126 neurons. Under these conditions, predictions are expected to be highly accurate (Extended Data Fig. 4a,e; correlation to ground truth:  $0.87 \pm 0.06$  for a noise level of 2, median  $\pm$  s.d.; Gaussian smoothing of the ground truth with  $\sigma=0.2$  s). Consistent with electrophysiological recordings<sup>44</sup>, spiking activity estimated by CASCADE with a model trained on DS #04 was sparse ( $0.6 \pm 1.1$  spikes during the initial 2.5 s of the odor response; mean  $\pm$  s.d.; Fig. 5a) and variable across neurons (Fig. 5b) and clearly different for the anatomically distinct dorsal and ventral regions of pDp ( $0.07 \pm 0.11$  Hz versus  $0.21 \pm 0.11$  Hz; entire recording).

The comparison of  $\Delta F/F$  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. 5c,d).

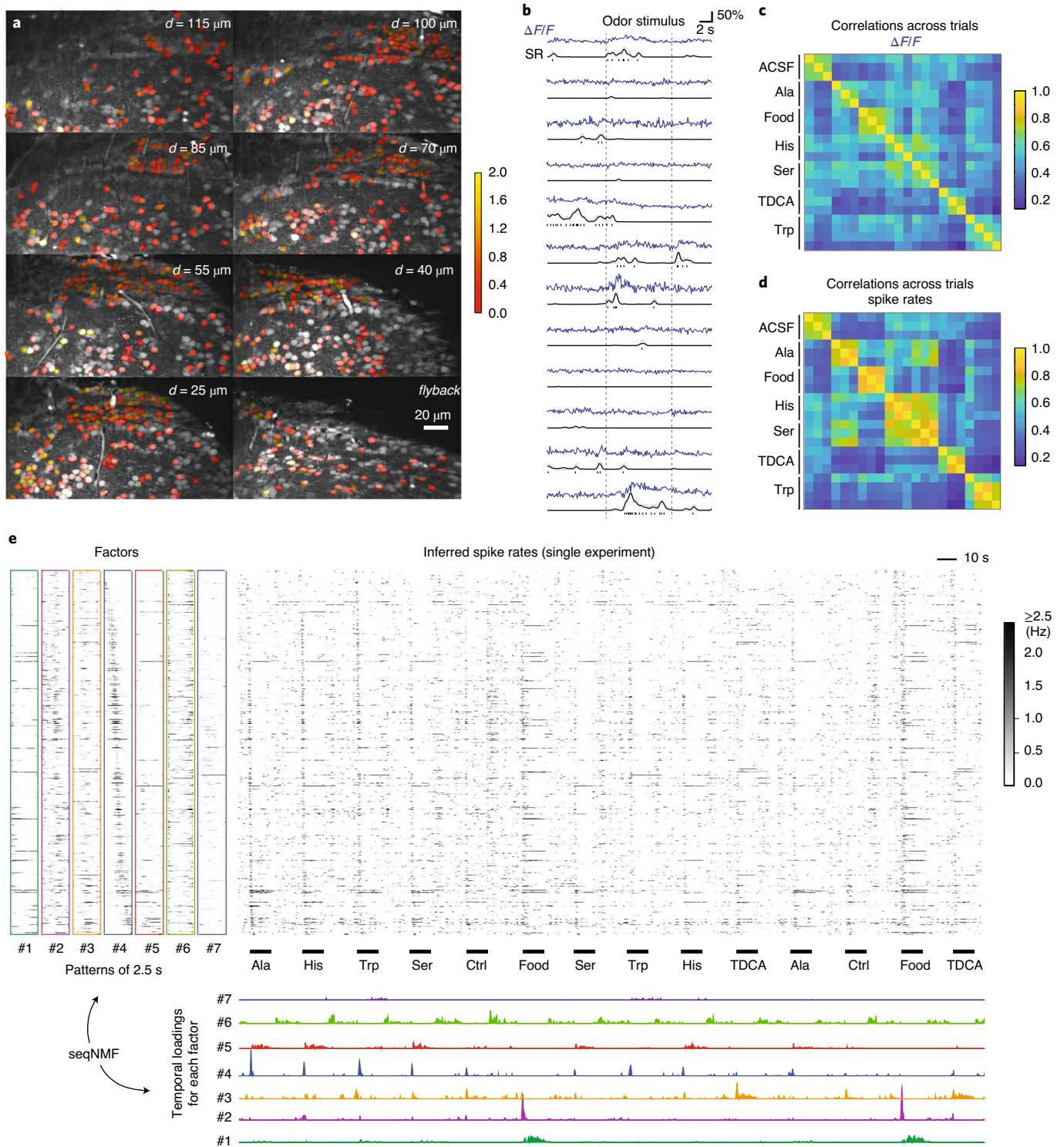
Previous studies showed that odor-evoked population activity in pDp is dynamic<sup>44,45</sup>, but the fine temporal structure has not been explored in detail. We analyzed inferred spike rate patterns using unsupervised non-negative matrix factorization for sequence

detection (seqNMF; ref. <sup>46</sup>) to identify recurring short (2.5-s) sequences of population activity (factors) in the overall population activity. Factors showed rich temporal structure on a sub-second time scale (Fig. 5e). Multiple factors were active with high precision and in a stimulus-specific manner at distinct phases of the odor response. For example, factors 2 and 4 in Fig. 5e were transient and associated with response onset; factor 5 persisted during odor presentation; and factor 6 was activated after stimulus offset (Fig. 5e). Odor-evoked population activity, therefore, exhibited complex dynamics on time scales that cannot be resolved without temporal deconvolution. The transformation of calcium signals into spike rate estimates by CASCADE, thus, provides interesting opportunities to use calcium imaging for the analysis of fast network dynamics.

We next analyzed the Allen Brain Observatory Visual Coding dataset, comprising more than 400 experiments in mice with transgenic GCaMP6f expression, each consisting of approximately 100–200 neurons recorded at very low noise levels ( $0.94 \pm 0.25$  %-Hz<sup>-1/2</sup>; mean  $\pm$  s.d.; Fig. 6a)<sup>47</sup>. Using the global EXC model of CASCADE, we estimated the absolute spike rates across all 38,466 neurons from different transgenic lines (Fig. 6b and Extended Data Fig. 10; Gaussian smoothing of the ground truth with  $\sigma=0.05$  s). Spike rates were well described by a log-normal distribution centered around 0.1–0.2 Hz (Fig. 6c). Given the sampling rate (30 Hz) and noise level of this dataset, we expect a correlation of  $0.89 \pm 0.18$ , an error of  $0.70 \pm 0.96$  and a bias of  $0.27 \pm 1.00$  (median  $\pm$  s.d. across neurons), based on our previous cross-dataset comparisons, which included transgenic lines used in this population imaging dataset (Fig. 3). Because generalization could not be tested across a large number of inhibitory neuron datasets (Fig. 3), we did not include interneuron experiments in our analysis. Inferred spike rates varied systematically across cortical layers, with highest activity in layer 5 (Fig. 6d,e). Inferred rates also varied across transgenic lines (Fig. 6d) and across stimuli presented, with highest activation during naturalistic stimuli (natural scenes or movies; Fig. 6e). These results provide a comprehensive description of neuronal activity in the mouse visual system and reveal systematic differences in neuronal activity across cell types, brain areas, cortical layers and stimuli.

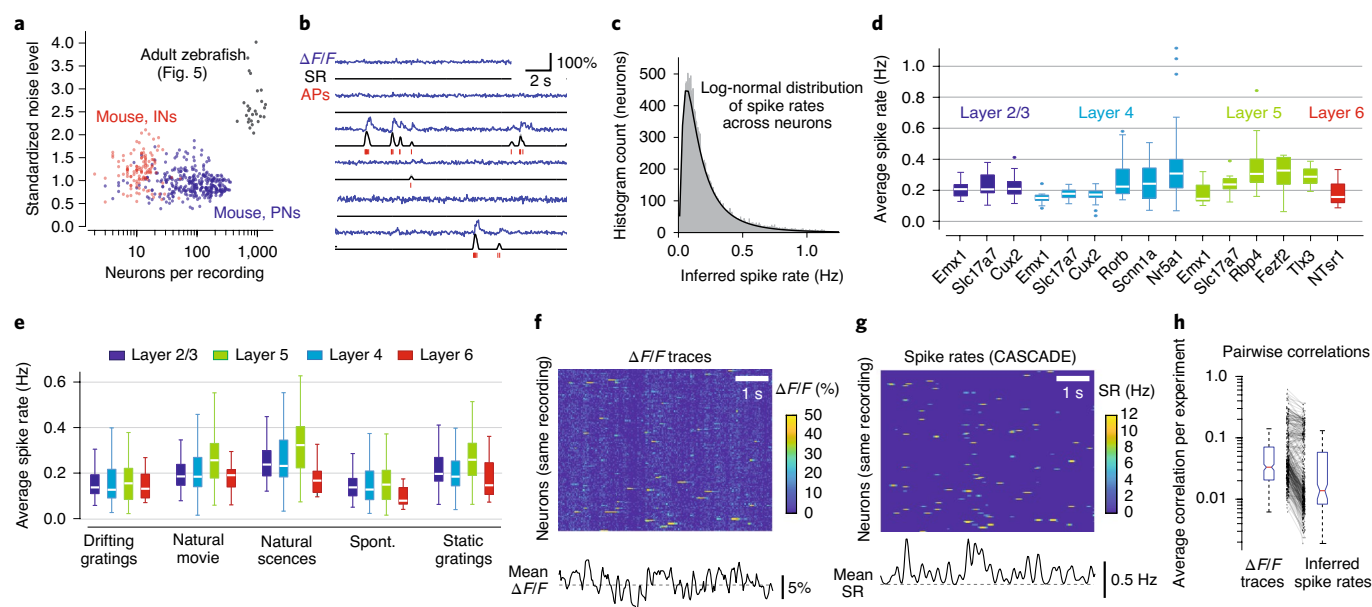
Raw  $\Delta F/F$  often exhibited correlated noise, visible as a vertical striping in matrix plots, which was small for individual neurons but tended to dominate the mean  $\Delta F/F$  across neurons, possibly due to technical noise or neuropil signal (Fig. 6f). CASCADE visibly eliminated these artifacts (Fig. 6g). As a consequence, correlations between activity traces of different neurons were reduced across all experiments by  $38\% \pm 43\%$  (mean  $\pm$  s.d.; Fig. 6h;  $P < 10^{-15}$ , paired signed-rank test). Using data simulated with NAOMi, we also found that spike inference by CASCADE brought measurements of pairwise firing rate correlations closer to the true values as compared to the raw calcium data (Supplementary Fig. 11). As many analyses of neuronal population activity require accurate measurements of pairwise neuronal correlations<sup>47,48</sup>, noise suppression and deconvolution by spike inference can help to make these analyses more reliable. These examples illustrate how calibrated spike inference by CASCADE can be applied to remove noise from calcium signals and to analyze the temporal structure of neuronal population dynamics.

**A user-friendly toolbox for spike inference.** The deployment of spike inference tools often creates practical problems. First, the difficulty to set up a computational pipeline might prevent widespread usage. We, therefore, generated a cloud-based solution using Colaboratory Notebooks that can be applied without local installations. We also set up a well-documented GitHub repository (<https://github.com/HelmchenLabSoftware/Cascade>) containing ground truth datasets, pre-trained models, notebooks and demo scripts that can be easily integrated into existing analysis pipelines, such as CalmAn, SIMA or Suite2P<sup>38,41,49</sup>. Because the algorithm works on regular laptops and workstations without GPU support, the main



**Fig. 5 | Inference of spiking activity with CASCADE from population calcium imaging across more than 1,100 neurons in adult zebrafish. a**, Multiple planes were imaged simultaneously. Similar results were obtained in 21 fish. The ROIs are colored with the average number of inferred stimulus-evoked spikes (color bar). Non-active neurons were left uncolored. **b**, Randomly selected examples of calcium traces ( $\Delta F/F$ , blue), inferred spike rates (black) and inferred discrete spikes, highlighting the de-noising through spike inference. **c**, Correlation of odor-evoked responses across trials, based on  $\Delta F/F$  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. SR, spike rate.





**Fig. 6 | Inference of spiking activity with CASCADE for the Allen Brain Observatory dataset in mice.** **a**, Number of recorded neurons versus standardized noise levels (in  $\% \cdot \text{Hz}^{-1/2}$ ) for all experiments from excitatory pyramidal neurons (PN, blue) and inhibitory neurons (IN, red); population imaging datasets in zebrafish (Fig. 5) 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 Extended Data Fig. 10 for more examples. **c**, Spike rates across the entire population are well described by a log-normal distribution (black fit).  $n = 38,466$  neurons. **d**, 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 ( $n = 336$  experiments). **e**, Average spike rates for different stimulus conditions ( $x$  labels) across layers (colors). Each data point is the mean spike rate across one experiment. **f**, Excerpt of raw  $\Delta F/F$  traces of a subset of neurons of a single experiment (L2/3-Slc17a7, experiment ID '652989705'). Correlated noise is visible as vertical striping patterns. **g**, Same as **f** but with inferred spike rates. **h**, Average correlation between neuron pairs within an experiment ( $n = 336$  experiments), computed from raw  $\Delta F/F$  traces (left) and inferred spike rates (right). For box plots, the median is indicated by the central line; 25th and 75th percentiles are indicated by the box; and maximum/minimum values excluding outliers (points) are indicated by the whiskers. AP, action potential; SR, spike rate.

installation difficulties of typical deep learning applications are circumvented.

In a typical workflow, the noise level for each neuron in a calcium imaging dataset is determined. Then, a model that has been pre-trained on noise-matched, resampled ground truth is loaded from an online library and applied to the  $\Delta F/F$  data without any need to adjust parameters. CASCADE can be easily modified and retrained to address further specific needs, such as more complex loss functions<sup>23</sup> or a modified architecture. Moreover, 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 (Supplementary Fig. 12).

A second problem is that experimenters might need additional tools and documentation for interpretation of the results. We, therefore, included graphical outputs and guiding comments that are accessible also for non-specialists throughout the demo scripts. Together with existing literature on the interpretation of raw calcium data<sup>4,5,23,40,50</sup>, these tools will help to focus the attention on data quality and make users aware of the potentials and limitations of raw and deconvolved data.

## Discussion

Any spike inference approach, in particular methods based on deep learning, critically depend on the availability and quality of ground truth data. We, therefore, created a ground truth database that is larger and more diverse than previous datasets<sup>16,19</sup> (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 invest primarily into the optimization of the inference algorithm itself. Unlike previous supervised spike inference algorithms<sup>16,19,20</sup>, CASCADE is not trained on fixed ground truth data but resamples the ground truth to match both frame rate and noise level automatically for each neuron (Fig. 2). This strategy significantly improved inference, highlighting the importance not only of realistic calcium signals but also of realistic noise patterns.

The generalization of spike inference methods across unseen datasets was investigated sporadically<sup>13,14,22</sup>, but systematic studies were lacking, presumably due to the scarcity of ground truth data. We, therefore, took advantage of our large database to explore how predictions depend on species (zebrafish or mouse), indicator type, brain region (Fig. 3) and other potentially important experimental parameters. Surprisingly, some training datasets allowed for efficient generalization across these parameters, and a combined training dataset achieved uniformly high performance across all test sets. This result was obtained for both excitatory and inhibitory neurons, although absolute spike rates of inhibitory neurons were underestimated. The 'global EXC model', therefore, exhibits efficient generalization and is well suited for practical applications of spike inference in unseen datasets. Interestingly, some datasets performed poorly as training sets, whereas 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 can 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.

In comparison to other approaches, spike predictions by CASCADE were more precise, as measured by correlation metrics, but also less biased toward underestimates of true spike rates (Fig. 4). We reason that reliable spike inference critically depends on the balance between spike detection and noise suppression. Although over-suppression appears to be advantageous in less expressive models, deep networks appear to afford less suppression because their high expressiveness allows for highly specific differentiation between signal and noise. CASCADE exploits this feature while keeping the network size small, which prevents overfitting. In theory, it is possible that other algorithms outperform CASCADE in regimes that are not covered by the ground truth database (for example, extremely low noise levels or tonically spiking neurons that are transiently inhibited<sup>51</sup>). Our results also indicate that enhancing the diversity of ground truth datasets can be more efficient than simply increasing dataset size to achieve further improvements in performance (Extended Data Fig. 6).

CASCADE was not sensitive to user-adjustable hyper-parameters or the class of the deep networks tested, which has two practical consequences. First, it seems more valuable to optimize the acquisition of more specific and diverse ground truth and the pre-processing of calcium data rather than to focus on improvements of the deep networks. Second, because hyper-parameters do not need to be adjusted by the user, the application of spike inference becomes simple in practice. Although some previous studies assumed that user-adjustable parameters in model-based algorithms increase the interpretability of the model<sup>11–14</sup>, we argue here that (1) biophysical model parameters are often ambiguous<sup>13</sup> 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 expected error of the results and a detailed documentation in the Colaboratory Notebook with help for interpretation.

Quantitative inference of spike rates is critical for the analysis of existing and future calcium imaging datasets<sup>4,5,23</sup>. The approach usually requires single-neuron resolution and is less well suited for signals from multiple neurons, such as endoscopic one-photon data with high background fluorescence, fiber photometry or wide-field imaging. Moreover,  $\Delta F/F$  can, in theory, report only spike rate changes. Nevertheless, we found that absolute spike rates can be inferred when the baseline activity is sufficiently sparse to enable the determination of the fluorescence baseline level  $F_0$ , which was the case in all datasets examined here (Figs. 5 and 6). The enhanced temporal resolution will be particularly useful for the analysis of neuronal activity during natural stimulus sequences and behaviors that occur on time scales shorter than typical durations of calcium transients, such as dynamical neuronal representations across theta cycles<sup>52</sup> or early and late phases of sensory responses in cortical areas<sup>53</sup>. Moreover, the inference of absolute spike rates will help improve the calibration of precisely patterned optogenetic manipulations<sup>54,55</sup> and the extraction of constraints—for example, 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. Future work should, thus, focus on the reduction of movement artifacts and neuropil contamination both by experimental design<sup>40,56</sup> and by extraction methods<sup>38,39</sup>, including the correct estimation of the  $F_0$  baseline despite unknown background fluorescence. In the long term, the development of more linear calcium indicators<sup>57</sup> and especially the acquisition and integration of more specific ground truth—for example, for additional interneurons and subcortical brain regions—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.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41593-021-00895-5>.

Received: 14 August 2020; Accepted: 23 June 2021;

Published online: 2 August 2021

### References

- Göbel, W. & Helmchen, F. In vivo calcium imaging of neural network function. *Physiology* **22**, 358–365 (2007).
- Harris, K. D., Quiroga, R. Q., Freeman, J. & Smith, S. L. Improving data quality in neuronal population recordings. *Nat. Neurosci.* **19**, 1165–1174 (2016).
- Rose, T., Goltstein, P. M., Portugues, R. & Griesbeck, O. Putting a finishing touch on GECIs. *Front. Mol. Neurosci.* **7**, 88 (2014).
- Sabatini, B. L. The impact of reporter kinetics on the interpretation of data gathered with fluorescent reporters. Preprint at <https://www.biorxiv.org/content/10.1101/834895v1> (2019).
- Wei, Z. et al. A comparison of neuronal population dynamics measured with calcium imaging and electrophysiology. *PLoS Comput. Biol.* **16**, e1008198 (2020).
- Ali, F. & Kwan, A. C. Interpreting in vivo calcium signals from neuronal cell bodies, axons, and dendrites: a review. *Neurophotonics* **7**, 011402 (2020).
- Yaksi, E. & Friedrich, R. W. Reconstruction of firing rate changes across neuronal populations by temporally deconvolved  $\text{Ca}^{2+}$  imaging. *Nat. Methods* **3**, 377–383 (2006).
- 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).
- Vogelstein, J. T. et al. Spike inference from calcium imaging using sequential Monte Carlo methods. *Biophys. J.* **97**, 636–655 (2009).
- Vogelstein, J. T. et al. Fast nonnegative deconvolution for spike train inference from population calcium imaging. *J. Neurophysiol.* **104**, 3691–3704 (2010).
- 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**, 201 (2013).
- 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).
- Greenberg, D. S. et al. Accurate action potential inference from a calcium sensor protein through biophysical modeling. Preprint at <https://www.biorxiv.org/content/10.1101/479055v1> (2018).
- Pachitariu, M., Stringer, C. & Harris, K. D. Robustness of spike deconvolution for neuronal calcium imaging. *J. Neurosci.* **38**, 7976–7985 (2018).
- Friedrich, J., Zhou, P. & Paninski, L. Fast online deconvolution of calcium imaging data. *PLoS Comput. Biol.* **13**, e1005423 (2017).
- Berens, P. et al. Community-based benchmarking improves spike rate inference from two-photon calcium imaging data. *PLoS Comput. Biol.* **14**, e1006157 (2018).
- Jewell, S. & Witten, D. Exact spike inference via  $l_0$  optimization. *Ann. Appl. Stat.* **12**, 2457–2482 (2018).
- 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).
- Theis, L. et al. Benchmarking spike rate inference in population calcium imaging. *Neuron* **90**, 471–482 (2016).
- Sebastian, J., Sur, M., Murthy, H. A. & Magimai-Doss, M. Signal-to-signal neural networks for improved spike estimation from calcium imaging data. *PLoS Comput. Biol.* **17**, e1007921 (2021).
- Hoang, H. et al. Improved hyperaccuracy estimation of spike timing from calcium imaging. *Sci. Rep.* **10**, 17844 (2020).
- Éltes, T., Szoboszlai, M., Kerti-Szigeti, K. & Nusser, Z. Improved spike inference accuracy by estimating the peak amplitude of unitary  $[\text{Ca}^{2+}]$  transients in weakly GCaMP6f-expressing hippocampal pyramidal cells. *J. Physiol.* **597**, 2925–2947 (2019).
- Evans, M. H., Petersen, R. S. & Humphries, M. D. On the use of calcium deconvolution algorithms in practical contexts. Preprint at <https://www.biorxiv.org/content/10.1101/871137v1> (2019).
- Zhu, P., Fajardo, O., Shum, J., Zhang Schäfer, 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).

25. Schoenfeld, G., Carta, S., Rupprecht, P., Ayaz, A. & Helmchen, F. In vivo calcium imaging of CA3 pyramidal neuron populations in adult mouse hippocampus. Preprint at <https://www.biorxiv.org/content/10.1101/2021.01.21.427642v1> (2021).
26. 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).
27. 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).
28. 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).
29. Kwan, A. C. & Dan, Y. Dissection of cortical microcircuits by single-neuron stimulation in vivo. *Curr. Biol. CB* **22**, 1459–1467 (2012).
30. Huang, L. et al. Relationship between simultaneously recorded spiking activity and fluorescence signal in GCaMP6 transgenic mice. *Elife* **10**, e51675 (2021).
31. Ledochowitsch, P. et al. On the correspondence of electrical and optical physiology in in vivo population-scale two-photon calcium imaging. Preprint at <https://www.biorxiv.org/content/10.1101/800102v1> (2019).
32. Dana, H. et al. Sensitive red protein calcium indicators for imaging neural activity. *eLife* **5**, e12727 (2016).
33. Akerboom, J. et al. Optimization of a GCaMP calcium indicator for neural activity imaging. *J. Neurosci.* **32**, 13819–13840 (2012).
34. Chen, T.-W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
35. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
36. Mathis, A. et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
37. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition* <https://ieeexplore.ieee.org/document/5206848> (2009).
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. 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. *J. Neurosci. Methods* **358**, 109173 (2019).
41. Pachitariu, M. et al. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. Preprint at <https://www.biorxiv.org/content/10.1101/061507v2> (2017).
42. Jewell, S., Hocking, T. D., Fearnhead, P. & Witten, D. Fast nonconvex deconvolution of calcium imaging data. *Biostatistics* **21**, 709–726 (2019).
43. 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).
44. Blumhagen, F. et al. Neuronal filtering of multiplexed odour representations. *Nature* **479**, 493–498 (2011).
45. Rupprecht, P. & Friedrich, R. W. Precise synaptic balance in the zebrafish homolog of olfactory cortex. *Neuron* **100**, 669–683.e5 (2018).
46. Mackevicius, E. L. et al. Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience. *eLife* **8**, e38471 (2019).
47. 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).
48. Lin, I.-C., Okun, M., Carandini, M. & Harris, K. D. The nature of shared cortical variability. *Neuron* **87**, 644–656 (2015).
49. Kaifosh, P., Zaremba, J. D., Danielson, N. B. & Losonczy, A. SIMA: Python software for analysis of dynamic fluorescence imaging data. *Front. Neuroinformatics* **8**, 80 (2014).
50. Siegle, J. H. et al. Reconciling functional differences in populations of neurons recorded with two-photon imaging and electrophysiology. Preprint at <https://www.biorxiv.org/content/10.1101/2020.08.10.244723v1.full> (2020).
51. Vanwalleghem, G., Constantin, L. & Scott, E. K. Calcium imaging and the curse of negativity. *Front. Neural Circuits* **14**, 607391 (2021).
52. Kay, K. et al. Constant sub-second cycling between representations of possible futures in the hippocampus. *Cell* **180**, 552–567 (2020).
53. van der Bourg, A. et al. Temporal refinement of sensory-evoked activity across layers in developing mouse barrel cortex. *Eur. J. Neurosci.* **50**, 2955–2969 (2019).
54. Pégard, N. C. et al. Three-dimensional scanless holographic optogenetics with temporal focusing (3D-SHOT). *Nat. Commun.* **8**, 1228 (2017).
55. Packer, A. M., Russell, L. E., Dalglish, 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. Griffiths, V. A. et al. Real-time 3D movement correction for two-photon imaging in behaving animals. *Nat. Methods* **17**, 741–748 (2020).
57. 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 (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

## Methods

**Ground truth recordings in adult zebrafish.** All zebrafish experiments were approved by the Veterinary Department of the Canton Basel-Stadt (Switzerland). For the recordings in DS #04 and DS #05, the adult zebrafish brain was dissected *ex vivo*<sup>24</sup>, and OGB-1 AM or Cal-520 AM were injected in pDp as described<sup>58</sup>. The dura mater above pDp was carefully removed to prevent clogging of the patch pipette. Calcium indicators were injected for 1–2 min at two locations (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, respectively) and was monitored by snapshot multi-photon images. The pressure was adjusted to avoid fast swelling of the tissue.

Juxtacellular recordings were performed >1 h and <4 h after the dye injection. Patch pipettes were pulled from 1-mm borosilicate glass capillaries (Hilgenberg) with a resistance of 5–8 MΩ and backfilled with artificial cerebrospinal fluid (ACSF) (in mM: 124 NaCl, 2 KCl, 1.25 KH<sub>2</sub>PO<sub>4</sub>, 1.6 MgSO<sub>4</sub>, 22 D-(+)-glucose, 2 CaCl<sub>2</sub>, 24 NaHCO<sub>3</sub>; pH 7.2; 300–310 mOsm) containing 0.05 mM Alexa Fluor 594.

The explant preparation was rotated about the anterior–posterior axis to allow for optical access from the side (sagittal imaging). Using a multi-photon microscope, images generated from fluorescence and from the asymmetry of the signal on a four-quadrant detector for transmitted light were used to target the pipette to pDp, while continuous low pressure (30–40 mbar) was applied to prevent clogging. 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<sup>45,59</sup> but with lower pressure. Juxtacellular recordings were performed after establishing a loose seal (typically 30–50 MΩ) with a target neuron. In some cases, a small negative pressure was applied initially to improve the electrical contact with the target cell. In several cases, micropipettes were reused multiple times. Recordings were performed in voltage-clamp mode with the voltage adjusted such that the resulting current approximated zero<sup>60</sup>.

For DSs #06–08, which were based on a transgenic line expressing GCaMP6f in the forebrain<sup>43</sup>, the experimental procedures were similar except for the injection of synthetic dyes. Because the baseline brightness of GCaMP6f is low, it was often difficult to identify individual neurons. Upon application of odor stimuli, stimulus-responsive neurons that expressed GCaMP6f became brighter, which permitted reliable visual identification for targeted patching. For regions in the dorsal telencephalon (DS #07) with no obvious odor responses, cells were patched randomly based on shadow images generated by the blown-out Alexa dye<sup>59</sup>.

Simultaneous recordings of fluorescence and extracellular spikes of the same neuron were synchronized using ScanImage 3.8 for imaging<sup>61</sup> and Ephys for electrophysiology<sup>62</sup>. Calcium imaging was performed at intermediate zoom (Fig. 1) with a frame rate of 7.5 Hz or 7.8125 Hz for DS #04 and DS #05 and at high zoom with a frame rate of 30 Hz for DSs #06–08. Electrophysiological recordings were low-pass filtered at 4 kHz (4-pole Bessel filter) and sampled at 10 kHz.

Recordings were performed in 120-s episodes, and food extract was applied to the nose as described<sup>45</sup>. In pDp, spike rates are usually very low. When no spiking activity was observed, the holding potential of the pipette was set to higher values (between +5 mV and +30 mV) to generate a depolarizing extracellular current that generated spikes if the seal resistance was sufficiently high. 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 in zebrafish ground truth datasets.** DS #04: 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<sup>58</sup>. Because OGB-1 localizes predominantly to the nucleus and because the resolution was high, neuropil contamination is negligible in this dataset.

DS #05: 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 labelled only a small central volume in pDp.

DS #06: 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 frame rate were chosen to match previous experiments<sup>45</sup>.

DS #07: tg(NeuroD:GCaMP6f), dorsal part of the dorsal pallium in adult zebrafish. All recorded neurons were mapped onto brain regions Dm, Dl, rDc and cDc based on neuroD expression in the dorsal part of the dorsal pallium (Supplementary Fig. 13, following Huang et al.<sup>63</sup>). Although this region is not known to be directly involved in olfactory processing, we noticed that several neurons were inhibited during odor stimulation (duration, 10–30 s).

DS #08: tg(NeuroD), olfactory bulb (OB) in adult zebrafish. In the OB of this transgenic line, GCaMP6f is restricted to a distinct, small subset of putative mitral cells and interneurons<sup>43</sup>. Neurons 1–3, 5 and 7 were identified as interneurons based on their small size and morphology, whereas neurons 4, 6, 8 and 9 were classified as putative mitral cells.

**Ground truth recordings in anaesthetized mice.** All experimental procedures related to DS #18 and DS #19 were approved by the Cantonal Veterinary Office

in Zurich (Switzerland). Mice were kept on a reversed 12-h light/dark cycle. For virus-induced expression of R-CaMP1.07 (DS #19), AAV1-EFα1-R-CaMP1.07 and AAV1-EFα1-DIO-R-CaMP1.07 were stereotactically injected under isoflurane anaesthesia into the barrel cortex of C57BL/6J mice and into hippocampal area CA3 of tg(Grik4-cre)G32-4Stl mice as described<sup>26</sup>. We combined electrophysiology and calcium imaging in acute experiments in anaesthetized animals (*n* = 3; at least 2 weeks after virus injection) as described<sup>26</sup>. A stainless steel plate was fixed to the exposed skull using dental acrylic cement. A 1 × 1 mm<sup>2</sup> craniotomy was made over the barrel cortex. The dura mater was cleaned with Ringer's solution (containing in mM: 135 NaCl, 5.4 KCl, 1.8 CaCl<sub>2</sub>, 5 HEPES, pH 7.2 with NaOH) and carefully removed. To reduce tissue motion caused by heartbeat and breathing, the craniotomy was filled with low-concentration agarose gel and gently pressed with a glass coverslip. For CA3 recordings (DS #18), a 4-mm Ø craniotomy was centred over the injection site. The overlying cortex was aspirated until the corpus callosum became visible. The cavity was filled with 1% agarose 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 Fluor 488 (Invitrogen) was added to the solution, or pipettes were coated with BSA Alexa Fluor 594 (Invitrogen). Action potentials were recorded in current-clamp using an Axoclamp 2B amplifier (Axon Instruments, Molecular Devices) and digitized at 10 kHz using Clampex 10.2 software. Calcium recordings were performed using HelioScan<sup>64</sup>.

The care of animals and experimental procedures related to DS #03 were carried out in accordance with national and institutional guidelines, and all experimental protocols were approved by the Animal Experimental Committee of the University of Tokyo. Mice were kept in a non-inverted 12-h light/dark cycle. Ambient temperature and humidity of the animal room were controlled at 20–25 °C and 40–60%, respectively. C57BL6/J male mice were anaesthetized by intraperitoneal injection of 1.9 mg g<sup>-1</sup> urethane, and the skull was partly exposed and attached to a stainless steel frame as described<sup>27</sup>. In a small craniotomy over the barrel cortex, we removed the dura, filled the cranial window with 1.5% agarose and placed a coverslip over the agarose to minimize brain movements<sup>27</sup>. 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<sup>27</sup>. 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 Fluor 594 (50 µM), inserted pipettes into the barrel and targeted Cal-520-loaded somata. Approximately 10 min after establishing the loose-seal cell-attached configuration, we performed simultaneous recordings and high-speed line-scan calcium imaging (500 Hz) on the soma of cortical neurons as described<sup>27</sup>. The electrophysiological data were filtered at 10 kHz and digitized at 20 kHz by using MultiClamp 700B and DigiData 1322A (Molecular Devices) and acquired using AxoGraph X (AxoGraph).

Experiments for DSs #24–27 were approved by the Veterinary Department of the Canton Basel-Stadt (Switzerland). Mice were kept on an inverted light cycle. DSs #24–26 were recorded in slices of mouse visual cortex as described<sup>28</sup>. Inhibitory neurons were targeted by injecting GCaMP6f-expressing AAV1 virus into PV-Cre, VIP-Cre or SOM-Cre mice. Coronal slices were cut with a thickness of 350 µm, and loose-patch recordings were performed at 32 °C in ACSF with WinWCP software (John Dempster). To induce activity in otherwise quiet slices, a potassium-based solution was applied to the slice through a second pipette. Simultaneous calcium imaging was performed with a two-photon microscope recording at 34 Hz through a ×16 water immersion objective (0.8 NA, Nikon)<sup>28</sup>. DS #27 was recorded in anaesthetized mice as described<sup>45</sup>. Adult (>8 weeks) PV-tTomato mice (cross between Rosa-CAG-LSL-tTomato (JAX: 007914) and PV-Cre (JAX: 008069)) were injected with GCaMP6f-AAV (AAV1.Syn.GCaMP6f.WPRE.SV40, UPENN) in primary visual cortex (V1, ~2.5 mm lateral, ~0.7 mm anterior of the posterior suture). Acute recordings were performed at least 2 weeks after the initial injection. Mice were initially anaesthetized with a mixture of fentanyl (0.05 mg ml<sup>-1</sup>), midazolam (5.0 mg kg<sup>-1</sup>) and medetomidin (0.5 mg kg<sup>-1</sup>); a metal headplate was fixed on the skull; and a craniotomy was made above V1. Anaesthesia was maintained with a low concentration of isoflurane (0.5% in O<sub>2</sub>). Borosilicate glass pipettes (6–8 MΩ) filled with a solution containing 110 mM potassium gluconate, 4 mM NaCl, 40 mM HEPES, 2 mM ATP-Mg, 0.3 mM GTP-NaCl and 0.03 mM Alexa Fluor 594 (adjusted to pH 7.2 with KOH, ~290 mOsm) were lowered into the visual cortex. Neurons expressing GCaMP6f and tTomato were targeted for juxtacellular recordings in loose-cell configuration under a two-photon microscope. For simultaneous electrophysiological and optical recordings, fluorescence was recorded with ScanImage<sup>61</sup> at 30 Hz, and juxtacellular voltage was recorded using a MultiClamp 700B amplifier (Axon Instruments). Signals recorded in slices were filtered at 1 kHz or 2 kHz, and signals recorded *in vivo* were filtered at 10 kHz before digitization at 20 kHz (National Instruments). 50-Hz noise was reduced by a noise eliminator (HumBug).

All experimental procedures for DS #02, DS #22 and DS #23 were performed in accordance with National Institutes of Health guidelines and approved by the Animal Care and Use Committee at the University of California, Berkeley. Mice were kept on a non-reversed 12-h light/dark cycle. These datasets were recorded in mouse primary visual cortex as described<sup>29</sup>. GFP-GIN mice were used to target



SOM interneurons; PV-Cre mice crossed with loxP-flanked tdTomato reporter mice were used to target PV interneurons; and CaMKII $\alpha$ -Cre mice crossed with loxP-flanked tdTomato mice were used to target excitatory neurons. One hour after loading of OGB-1 into V1 (ref. <sup>29</sup>), two-photon microscopy was used to target neurons 150–300  $\mu\text{m}$  below the brain surface with the recording pipette, while the mouse was anaesthetized with intraperitoneal injection of urethane and chlorprothixene. Two-photon imaging of neurons was performed with a  $\times 40$  objective at a frame rate of 15.6 Hz, while voltage was recorded in a loose-cell configuration from the same neuron as described<sup>29</sup>.

**Analysis of ground truth recordings.** Movies of calcium indicator fluorescence were corrected offline for movement artifacts (slow drifts due to relaxation of the brain tissue for zebrafish data; fast movements for recordings in anaesthetized mice). Ground truth recordings from DS #03 were not corrected for movement artifacts due to the scanning modality (line-scan). Thereafter, ROIs were manually drawn using a custom-written software tool (<https://git.io/vAeKZ>)<sup>45</sup> 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 (Supplementary Fig. 2).

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 manually adjusted threshold for each neuron extracted action potential events. The process of first generating a template that was afterwards used to detect stereotypic signals increased the signal-to-noise of detected events, similarly to previous usages of template matching in electrophysiology<sup>66,67</sup>.

**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. Calcium recordings with excessive movement artifacts or apparent inconsistencies of juxtacellular and calcium recordings were discarded entirely. Excessive movement artifacts were defined as events when the neuron visibly moved out of the imaging plane, such that transients generated by these movements were almost as frequent and prominent as true calcium transients. Apparent inconsistencies of recordings were identified as recordings where no spike events corresponded to visible calcium transients and where a spike-triggered average (Extended Data Fig. 1) did not show any signal, indicating that juxtacellular and calcium recordings were performed from different neurons. In addition, neurons were discarded when they did not spike at all even after application of currents or when they became visibly brighter after establishing a loose seal due to unknown, possibly mechanical, reasons. When the calcium recording clearly contained events without corresponding electrophysiological action potentials, the calcium trace of the manually drawn ROI and the calcium traces of adjacent neurons or neuropil were inspected together with the electrophysiological recordings to assess optical bleed-through, and ROIs were adjusted if necessary to avoid contamination. Occasionally, we also noted that mechanical stress exerted by the recording pipette can increase the brightness of the recorded neuron<sup>31</sup>, possibly by the release of calcium from internal stores. Recordings made during and after such events were discarded. Bursting can lead to adaptation of the extracellularly measured spike amplitude. Such recordings (for example, in DS #18 with bursts of more than ten action potentials with an inter-spike interval of approximately 5 ms) were carefully inspected for missed low-amplitude action potentials, in particular during these bursts.

**Extraction of ground truth from publicly available datasets.** Additional ground truth was extracted from publicly available datasets and quality controlled for each neuron<sup>16,19,30,32–34</sup>.

**The Allen Institute datasets.** For DSs #10–13 from ref. <sup>30</sup>, raw fluorescence traces were extracted from the processed datasets, which were downloaded from <https://portal.brain-map.org/explore/circuits/oephys>. Neuropil signal 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<sup>30</sup>. 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 as described in the Methods subsection ‘Quality control’.

**The Spikefinder datasets.** For DS #01, DS #15 and DS #16 from ref. <sup>19</sup>, the ground truth recordings at their native sampling rates as released during the Spikefinder challenge<sup>16</sup> were processed. This Spikefinder dataset consists of five separate

datasets. Spikefinder datasets #1 and #4 were excluded because fluorescence baseline and scaling were unknown. The other datasets were extracted as fluorescence traces, and  $F_0$  was computed as the 10th percentile value (adjusted depending on the spike rate of each neuron) and used to compute  $\Delta F/F_0$ . Some ground truth neurons were discarded due to a highly unstable fluorescence baseline, but no strict quality control was possible because the raw calcium imaging data were not available. As found previously, some datasets of the Spikefinder challenge come with calcium recordings that are delayed with respect to the electrophysiological recordings<sup>16</sup>. 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.** DS #09, DS #014, DS #017, DS #20 and DS #21 were downloaded from <http://crcns.org/data-sets/methods><sup>32–34,68,69</sup>. For DS #09 and DS #14 (ref. <sup>34</sup>), 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 signal was subtracted using the same standard scaling value for all neurons (neuropil contamination ratio, 0.7)<sup>34</sup>.  $F_0$  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 DS #17, DS #20 and DS #21, no raw calcium imaging data were available, therefore not allowing for strict quality control using raw calcium recordings as additional feedback. Neuropil signal was subtracted from raw fluorescence using the same standard scaling value for all neurons (neuropil contamination ratio, 0.7)<sup>32,33</sup>.  $F_0$  values were computed using percentile values that were adjusted to the noisiness of the recording and over window durations that were adjusted to baseline activity.

**Population calcium imaging with OGB-1 in zebrafish pDp.** Ex vivo preparations, OGB-1 AM injections and calcium imaging were performed as described for juxtacellular recordings. Calcium imaging was performed using a custom-built, multi-plane, multi-photon microscope based on a voice coil motor for fast z-scanning<sup>42</sup>. 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 in Dp was performed in eight planes (256  $\times$  512 pixels, approximately 100  $\times$  200  $\mu\text{m}$ ) at 7.5 Hz over a z-range of approximately 100  $\mu\text{m}$ . Due to slowly relaxing brain tissue, movement correction was applied every 5 min by acquiring local z-stacks with a z-range of  $\pm 6 \mu\text{m}$ . 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 and Trp, Sigma-Aldrich) were diluted to a final concentration of  $10^{-4}$  M, and bile acid (TDCA, Sigma-Aldrich) was diluted to  $10^{-5}$  M in ACSF immediately before the experiment. Food extract was prepared as described<sup>45</sup>. Odors were applied for 10 s through a constant stream of ACSF using a computer-controlled peristaltic pump<sup>45</sup> in pseudo-random order with three repetitions of each odor presentation.

**Extraction of linear kernels from ground truth data.** Linear kernels were extracted by 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.

To compute the variability of linear kernels across neurons within and across datasets (Extended Data Fig. 1), we split the ground truth recording of each neuron in five separate parts and computed the linear kernels for each of the segments separately. When the coefficient of variation across these five values was less than 0.5, the kernel amplitude was considered reliable and included in the plots (Extended Data Fig. 1).

**Computation of noise levels.** In the shot noise limited case, the mean fluorescence  $F_0$  scales with  $N$ , which is the number of photons collected by the detector per second, and the fluorescence baseline fluctuations  $\sigma_r$  scale with  $\sqrt{N}$ . Thus, the  $\Delta F/F$  baseline noise  $\sigma_{\Delta F/F} = \sigma_r/F_0$  scales with  $1/\sqrt{N}$ . If the fluorescence signal is sampled at frame rate  $f_r$  and the number of photons collected per frame reduces to  $N/f_r$ , thus  $\sigma_{\Delta F/F}$  scales with  $\sqrt{f_r}$ . To define a noise measure that is independent of frame rate, we, therefore, normalized  $\sigma_{\Delta F/F}$  for this shot noise effect and defined the standardized noise  $\nu$  as:

$$\nu = \frac{\sigma_{\Delta F/F}}{\sqrt{f_r}} = \frac{\text{Median}_t |\Delta F/F_{t+1} - \Delta F/F_t|}{\sqrt{f_r}} \quad (1)$$

The units for  $\nu$  are  $\% \cdot \text{Hz}^{-1/2}$ , which, for the purpose of readability, we omit in the text. When computed for  $\Delta F/F$  data in this way,  $\nu$  is quantitatively comparable across datasets. A value of  $\nu = 1$  indicates a very low noise level, whereas  $\nu = 8$  indicates a high noise level, independent of frame rate.



**Metrics to quantify performance of spike inference.** The ground truth spike rates were generated from discrete spikes by convolution with a Gaussian smoothing kernel (except in Supplementary Fig. 12, 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 ( $\sigma = 0.2$  s for 7.5-Hz recordings and  $\sigma = 0.05$  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<sup>16</sup> but does not contain information about absolute scaling or offsets.  $F_1$  scores combine false positives and false negatives<sup>12</sup> 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<sup>70</sup> but are less intuitive.

We defined three intuitive and complementary metrics (illustrated as color-coded equations in Supplementary Fig. 3). First, we used Pearson's correlation between ground truth spike rates and inferred spike rates as a standard measure of the similarity. Second, 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 incorrectly inferred or omitted spikes is about 70% of the number of spikes in the ground truth. Third, the (relative) 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 toward the minimum of the bias,  $-1$ , whereas other algorithms might tend to systematically overestimate spike occurrences (bias  $> 0$ ). Importantly, the error can be very high when the number of false positives and false negatives is high, but the bias might still be zero. Error and bias are, therefore, two metrics that describe the absolute errors in terms of spike rates, complementing the correlation metric.

**Architecture of the default convolutional network.** The default network consists of a standard convolutional network with six hidden layers, including three 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 and 5 time points), with an increasing number of features (20, 30 and 40 filters per layer). After the second and third layer, maximum pooling layers are inserted. A final densely connected hidden layer consisting of ten neurons relays the result to a single output neuron. Although all neurons in hidden layers are based on 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 to facilitate gradient descent with a time-symmetric Gaussian kernel of standard deviation 0.2 s for resampling at 7.5 Hz (unless otherwise indicated), and 0.05 s for 30 Hz, or with a causal kernel (inverse Gaussian distribution).

**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 optimizes not only the similarity of both signals (correlation) but also the absolute magnitude of the inferred spike rates. Based on errors computed via back-propagation, gradient descent was performed using a standard optimizer (adagrad; Supplementary Fig. 4). 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 10–20 epochs (except when analyzing overfitting; Supplementary Figs. 4 and 5).

In all spike inferences presented here, without 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 crucial and strictly distinct from the process of fitting parameters for a neuron or a dataset, which would yield better results for a given neuron but would fail to generalize to new data.

**Architecture of alternative deep learning networks.** All deep learning architectures (Supplementary Fig. 5) were trained with the same loss function, the same input and the same optimizer as the default network.

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

**Single convolutional layer network:** the first convolutional layer of the default network, a single max pooling layer and a single dense layer of ten neurons. Total of 1,021 trainable parameters.

**Deeper convolutional network (five convolutional neural network (CNN) layers):** five convolutional layers with filter sizes (11, 9, 7, 5 and 3) and filter

numbers (20, 30, 40, 40 and 40), three max pooling layers after the second, fourth and fifth convolutional layers and a final dense layer expansion of ten neurons. The reduction of the filter sizes compared to the default network is necessary because no zero-padding was applied, resulting in a decrease of the size of the one-dimensional trace with increasing layer depth. Total of 27,421 trainable parameters.

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

**Batch normalization:** same as the default network with batch normalization<sup>71</sup> 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), whereas 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.

**Naive 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 back-propagating through time<sup>72,73</sup>. 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<sup>74,75</sup>. 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 ReLUs. The network is, therefore, entirely linear but based on the same architecture (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 applied. 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 made 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. This approach is relatively slow but 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 Supplementary Fig. 7 and to allow for comparison against raw inferred spike rates, the resulting discrete spikes were convolved with the Gaussian smoothing kernel that was used to generate the ground truth. We provide a demo script that infers discrete spikes from spike rates predicted with CASCADE (available on GitHub: <https://git.io/JtZe4>).

**Generalized linear model to fit predictability across datasets.** To predict how well a model trained on a given ground truth dataset (for example, DS #08) is able to infer activity for another dataset (for example, DS #14), 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 (Supplementary Fig. 9). In total, eight predictors were used, separately or together.

First, indicator species was set to 1 if the training and test datasets had the same indicator species (synthetic dyes versus genetically encoded dyes) and 0 otherwise. Animal species was set to 1 if the training and test datasets had the same animal species (zebrafish versus mouse) and 0 otherwise. Spike rate was computed as the absolute difference between median spike rates across neurons from the 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 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<sup>76</sup>. 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 #18, 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 time course was computed as the correlation 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.

**Artificial ground truth generated with NAOMi.** The package NAOMi was used to generate simulated two-photon calcium recordings of neurons with known spike patterns<sup>40</sup>. These simulated datasets were used as ground truth to train CASCADE (Fig. 3) but also to test the effect of spike inference with CASCADE on the estimated pairwise correlations between neurons (Supplementary Fig. 11). We used the default parameters, which had been optimized for the simulation of GCaMP6f based on previous calibrations<sup>34</sup>. Artificial ground truth was generated at 30 Hz with a detection NA of 0.6 and an excitation NA of 0.8 at a depth of 100  $\mu\text{m}$  below the cortical surface, in a volume of  $250 \times 250 \times 100 \mu\text{m}^3$ . To increase the signal-to-noise ratio of the simulated recordings, we used a relatively high simulated laser power of 70 mW. We simulated recordings of the central plane of five such volumes over a duration of 166 s. We extracted the cleanest components of each simulation by selecting the spatial components (from the ideal components returned by NAOMi) that correlated most highly with the known ground truth signals (correlation with a known somatic ground truth signal  $>0.80$ ). We chose to only include the best-matching components because other components typically had much stronger neuropil contamination than our experimentally obtained ground truth recordings. Then, we extracted the fluorescence of the selected components and performed neuropil subtraction with a 2-pixel ring around the detected component using a factor of 0.45 for neuropil subtraction. Afterwards, we computed the  $\Delta F/F$  signal, using the 2nd percentile across the entire recording to determine  $F_0$ . This procedure resulted in ground truth recordings from a total of 250 simulated neurons.

**Adaptation of model-based spike inference algorithms.** The MLSpike algorithm<sup>12</sup> was downloaded from <https://github.com/MLSpikes/spikes> and used in MATLAB 2017a. 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 (DSs #01–05 and DS #22 and DS #23), a saturating non-linearity (saturation = 0.01) was used, whereas, for all other datasets, a GCaMP-like non-linearity (`pnonlin = (1.0 0.0)`) was defined and kept the same across datasets, because predictions have been described to depend only slightly on the precise values of the non-linearity<sup>12</sup>. 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 and all noise levels separately. The grid search ranged from 0.1 s to 5 s for tau and from 0.01 to 0.35 for amplitude.

The Peeling algorithm<sup>11</sup> was downloaded from <https://github.com/HelmchenLab/CalciumSim> and used in MATLAB 2017a. 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 s to 5 s for tau1 and from 2.5 to 35 for amp1. Discrete spike predictions were convolved with a Gaussian kernel such that the resulting trace optimized the loss function (mean squared error between predictions and ground truth).

The Python implementation of the  $L_1$ -regularized OASIS algorithm in CalmAn<sup>15</sup> was downloaded from <https://github.com/j-friedrich/OASIS> and used in Python 3.7. The constrained version of OASIS was used to reduce the number of free parameters, with only one single free parameter,  $g$ , that relates to the exponential time fluorescence decay constant  $\tau$  with the frame rate  $f$ :  $g = e^{-1/\tau f}$ . Grid search was performed for  $g$  in the range between 0.02 and 0.98, with a granularity of 0.02.

The Python implementation of the FastL0SpikeInference algorithm<sup>12</sup> (Jewell&Witten) was downloaded from <https://github.com/jewellsean/FastLZeroSpikeInference> and used in Python 3.7. A grid search was performed over two parameters for all ground truth datasets. Optimization was performed between 0.10 and 0.95 for the decay constant parameter gamma and between 0.0001 and 0.75 for the  $L_0$  parameter penalty. Discrete spike predictions were convolved with a Gaussian kernel such that the resulting trace optimized the loss function (mean squared error between predictions and ground truth).

The Python implementation of the OASIS algorithm in Suite2p<sup>41</sup> was downloaded from <https://github.com/MouseLand/suite2p> and used in Python 3.7.

Of three tunable parameters (tau, sig\_baseline and win\_baseline), only the first two significantly affected the performance of the algorithm in our hands. win\_baseline was set to 150 for all analyses. A grid search was performed over the two remaining parameters for all ground truth datasets. Optimization was performed between 0.5 and 3 for the decay time constant parameter tau and between 2.5 and 20 for the parameter sig\_baseline.

The optimal parameters resulting from the grid searches, which optimized the mean squared error between ground truth and inferred spike rates, are listed in Supplementary Table 1 and provided via GitHub (<https://git.io/tZe0>). In addition to these parameters, we further used Gaussian smoothing kernels of variable standard deviation to find the amount of smoothing for each algorithm and dataset that optimized the mean squared error. Finally, to compensate for the propensity of several model-based algorithms to infer spike rates with a temporal lag compared to ground truth spike rates, we tested time shifts between  $-1$  s and  $+1$  s and used the value that optimized the mean squared error for a given dataset to evaluate the algorithm in our analyses.

**Computational cost of spike inference.** The six investigated algorithms exhibit different behaviors 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, whereas CASCADE shows the opposite behavior due to its capability to parallelize spike inference. Therefore, all 26 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 10 s 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 to generate a 64-point-wide segment 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. 5 in an unsupervised fashion<sup>46</sup>. Based on initial parameter exploration, we used the following settings:  $K=7$ ,  $L=20$  and  $\lambda=0.002$ .  $K$  indicates the number of extracted patterns,  $L$  indicates the number of time points for each pattern and  $\lambda$  serves as a regularizer to decorrelate the detected patterns<sup>46</sup>. The results of this unsupervised non-negative matrix factorization approach are  $K=7$  temporal patterns that are each of them associated with a temporal loading that 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. 5).

**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<sup>47</sup>. 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. Because all recordings were performed at an imaging rate of approximately 30 Hz, a single set of CASCADE models ('global EXC model' at 30 Hz; Fig. 3a) 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 (for example, comparison across datasets), and the Wilcoxon signed-rank test was used for paired samples (for example, comparison of predictions for the same set of neurons using two different algorithms). Two-sided tests were applied unless noted otherwise. Effect sizes  $\Delta \pm$  confidence interval (CI) (pseudo-median  $\Delta$  and 95% CIs unless otherwise indicated) were computed in R. Box plots used standard settings in MATLAB, with the central line at the median of the distribution, the box at the 25th and 75th percentiles and the whiskers at extreme values excluding outliers (outliers defined as data points that are more than 1.5- $D$  away from the 25th or 75th percentile value, with  $D$  being the distance between the 25th and 75th percentiles).

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Ground truth data, including extracted spike times and calcium traces, are deposited in the GitHub repository together with demo scripts (<https://github.com/HelmchenLabSoftware/Cascade>). We provide a cloud-based Collaboratory Notebook that allows for interactive browsing through all datasets ([https://colab.research.google.com/github/HelmchenLabSoftware/Cascade/blob/master/Demo%20scripts/Explore\\_ground\\_truth\\_datasets.ipynb](https://colab.research.google.com/github/HelmchenLabSoftware/Cascade/blob/master/Demo%20scripts/Explore_ground_truth_datasets.ipynb)). Raw data were recorded in different formats, and all newly recorded raw datasets are also available upon reasonable request in their original formats. Publicly available datasets are described in detail in the Methods ('Extraction of ground truth from publicly available datasets').

Additional information on experimental design and reagents is available in the Research Life Sciences Reporting Summary linked to this paper.

### Code availability

A cloud-based version of CASCADE is available as a Colaboratory Notebook ([https://colab.research.google.com/github/HelmchenLabSoftware/Cascade/blob/master/Demo%20scripts/Calibrated\\_spike\\_inference\\_with\\_Cascade.ipynb](https://colab.research.google.com/github/HelmchenLabSoftware/Cascade/blob/master/Demo%20scripts/Calibrated_spike_inference_with_Cascade.ipynb)). The code is also available as a GitHub repository together with demo scripts, installation instructions and FAQs (<https://github.com/HelmchenLabSoftware/Cascade>). Pre-trained models for CASCADE are archived in an online server (<https://www.switch.ch/drive/>) and retrieved automatically by the CASCADE code.

### References

58. 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).
59. 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).
60. Perkins, K. L. Cell-attached voltage-clamp and current-clamp recording and stimulation techniques in brain slices. *J. Neurosci. Methods* **154**, 1–18 (2006).
61. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2**, 13 (2003).
62. Suter, B. A. et al. Ephus: multipurpose data acquisition software for neuroscience experiments. *Front. Neural Circuits* **4**, 100 (2010).
63. Huang, K.-H. et al. A virtual reality system to analyze neural activity and behavior in adult zebrafish. *Nat. Methods* **17**, 343–351 (2020).
64. Langer, D. et al. HelioScan: a software framework for controlling in vivo microscopy setups with high hardware flexibility, functional diversity and extendibility. *J. Neurosci. Methods* **215**, 38–52 (2013).
65. Pecka, M., Han, Y., Sader, E. & Mrsic-Flogel, T. D. Experience-dependent specialization of receptive field surround for selective coding of natural scenes. *Neuron* **84**, 457–469 (2014).
66. 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).
67. Guzman, S. J., Schlögl, A. & Schmidt-Hieber, C. Stimfit: quantifying electrophysiological data with Python. *Front. Neuroinformatics* **8**, 16 (2014).
68. GENIE project, Janelia Farm Campus, HHMI & Svoboda, K. Simultaneous imaging and loose-seal cell-attached electrical recordings from neurons expressing a variety of genetically encoded calcium indicators. <https://crcns.org/data-sets/methods/cai-1/about-cai-1> (2015).
69. Boaz, M., Dana, H., Kim, D. S., Svoboda, K. & GENIE project, Janelia Farm Campus, 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. <https://crcns.org/data-sets/methods/cai-2/about-cai-2> (2016).
70. 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).
71. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. 448–456 (PMLR, 2015).
72. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
73. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: continual prediction with LSTM. *Neural Comput.* **12**, 2451–2471 (1999).
74. Schuster, M. & Paliwal, K. Bidirectional recurrent neural networks. *Signal Process. IEEE Trans.* **45**, 2673–2681 (1997).
75. Graves A., Fernández S., Schmidhuber J. Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. Duch W., Kacprzyk J., Oja E., Zadrozny S. (eds) *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005. Lecture Notes in Computer Science*, vol 3697. (Springer, Berlin, Heidelberg, 2005).
76. Eden, U. T. & Kramer, M. A. Drawing inferences from Fano factor calculations. *J. Neurosci. Methods* **190**, 149–152 (2010).

### 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 P. Berens and E. Froudarakis for providing additional information on the Spikefinder datasets. We thank G. Schoenfeld for helpful discussions on DS #18 and H. Heiser, N. Temiz, C. Satou, G. Schoenfeld and H. 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 no. 310030-127091 and Sinergia grant no. CRSII5-18O316) and the European Research Council (ERC Advanced Grant BRAINCOMPAT, 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 no. 310030B-152833/1) and the European Research Council (ERC Advanced Grant MCircuits, grant agreement no. 742576); by the Novartis Research Foundation; by a UZH Forschungskredit and a fellowship from the Boehringer Ingelheim Fonds to P.R.

### Author contributions

P.R. conceived the project, developed the algorithm, performed ground truth recordings (DSs #4–8), performed all analyses, developed the toolbox and wrote the paper. S.C. performed ground truth recordings (DS #18 and DS #19). A.H. developed the toolbox. M.E. and K.K. (DS #3), A.K. and Y.D. (DS #2, DS #22 and DS #23) and A.B. and S.H. (DSs #24–27) performed and pre-processed ground truth recordings. F.H. supervised ground truth recordings (DS #18 and DS #19) and the development of the toolbox and wrote the paper. R.W.F. supervised ground truth recordings (DSs #4–8) and the development of the algorithm and wrote the paper.

### Competing interests

The authors declare no competing financial interests.

### Additional information

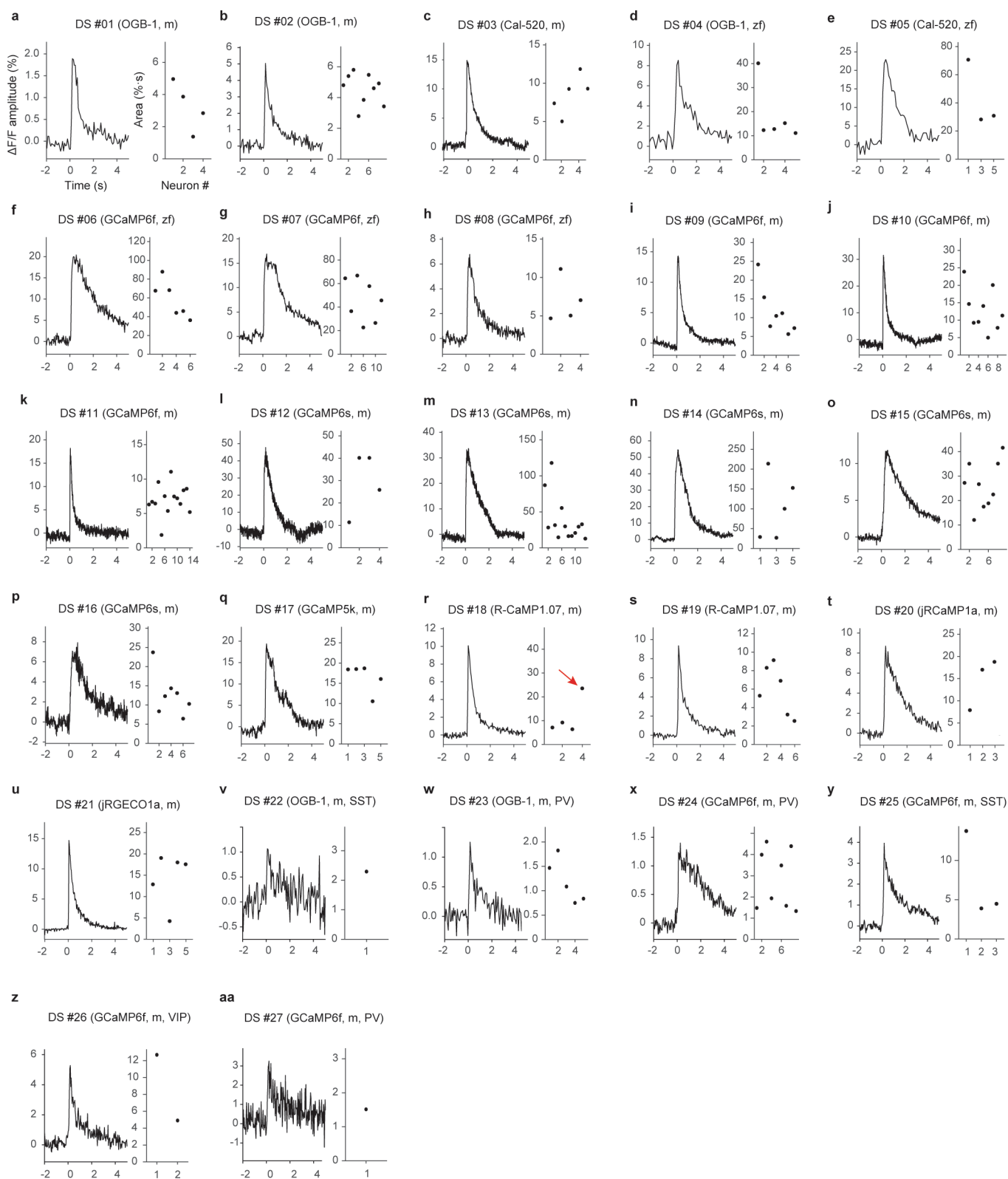
**Extended data** is available for this paper at <https://doi.org/10.1038/s41593-021-00895-5>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41593-021-00895-5>.

**Correspondence and requests for materials** should be addressed to P.R., F.H. or R.W.F.

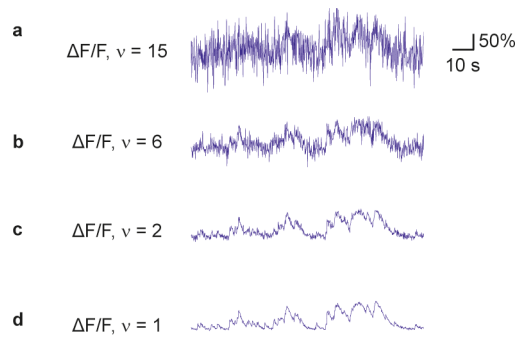
**Peer review information** *Nature Neuroscience* thanks the anonymous reviewers for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

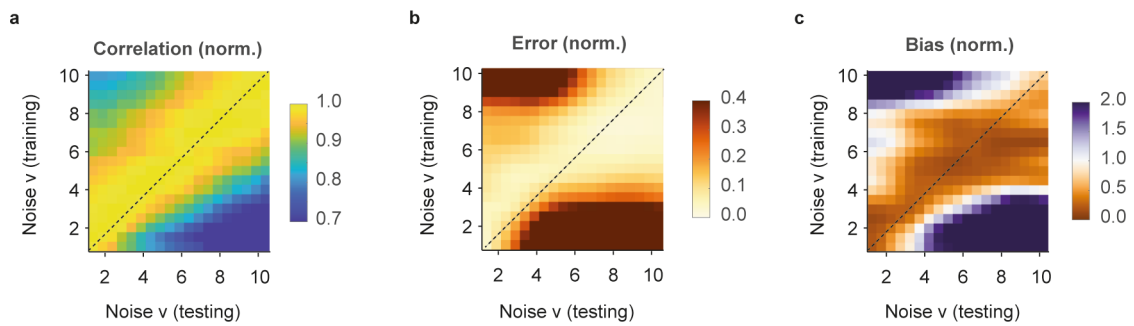


**Extended Data Fig. 1 | Linear kernels extracted from all ground truth datasets.** The kernels are optimized such that when the ground truth spike times are linearly convolved with the kernel, the experimentally recorded  $\Delta F/F$  trace is ideally approximated. In practice, this is achieved using 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) is 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 (r) indicates an outlier case that is discussed in Extended Data Fig. 4a. m: Mouse, zf: Zebrafish.

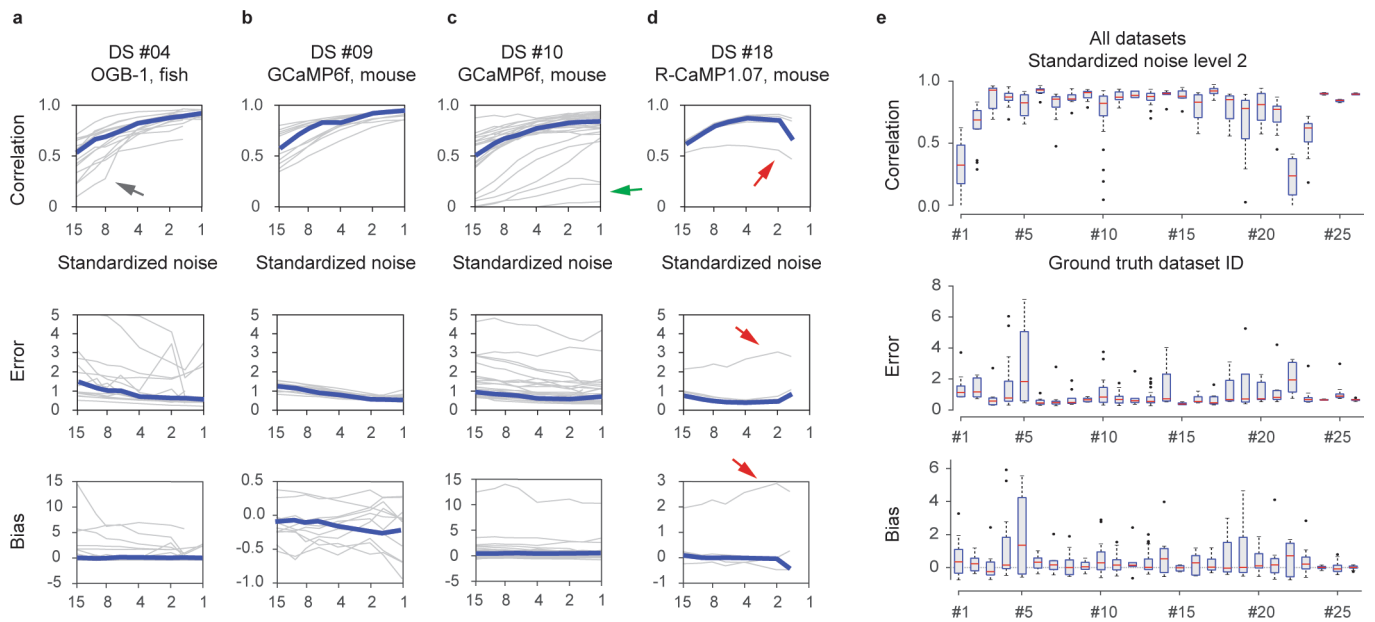




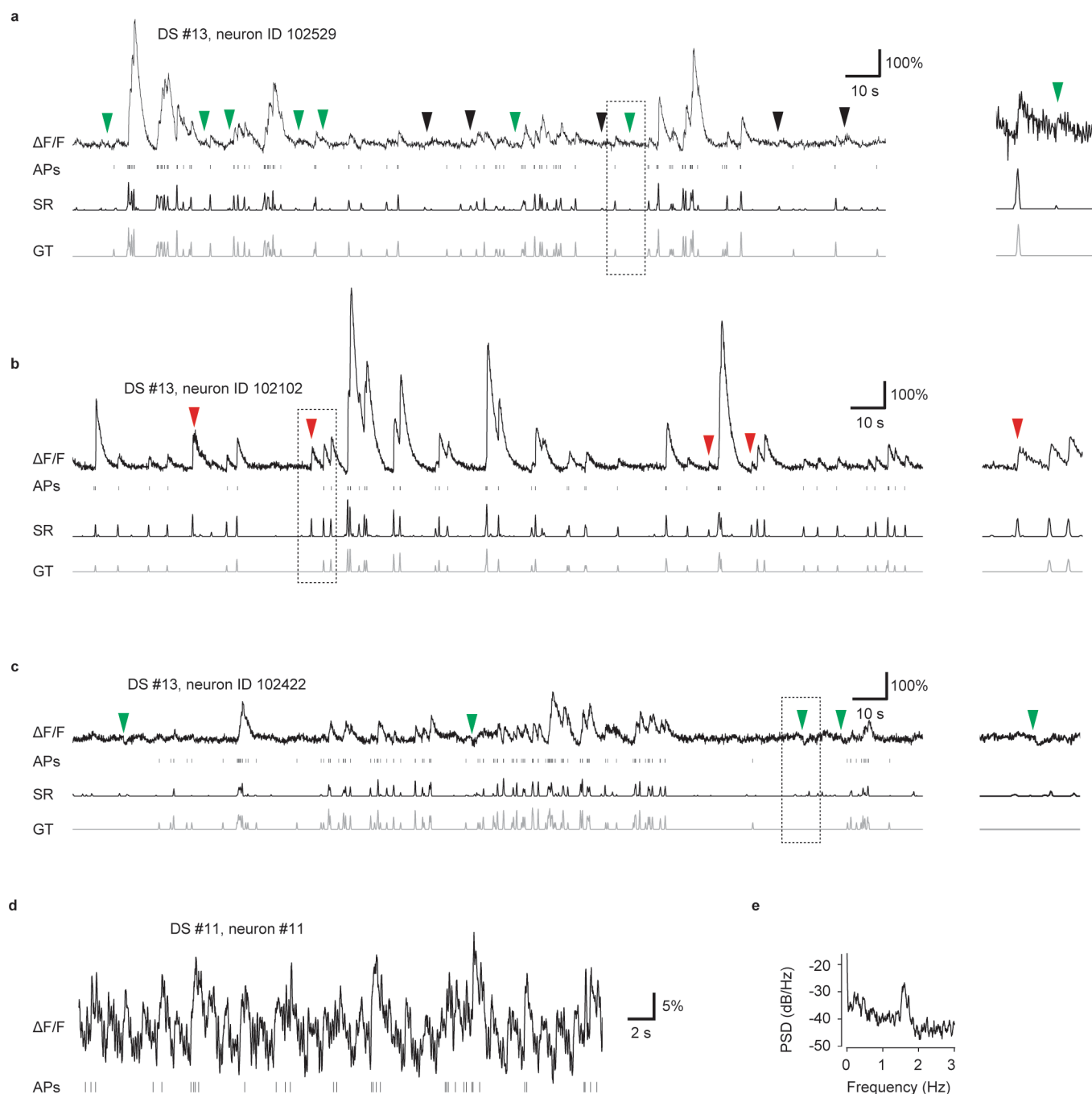
**Extended Data Fig. 2 | Illustration of different baseline noise levels.**  $\Delta F/F$  ground truth traces were resampled with added noise to reach the target noise level  $\nu$ . **a-d**, Noise level illustration from  $\nu=15$  (very high noise level) to  $\nu=1$  (very low noise level). Standardized noise  $\nu$  is given in units of  $\% \cdot \text{Hz}^{-1/2}$ .



**Extended Data Fig. 3 | Matching standardized noise level  $\nu$  of training and test data.** Same as Fig. 2e-g, 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 of each column.  $\nu$  in units of standardized noise,  $\% \cdot \text{Hz}^{-1/2}$ .

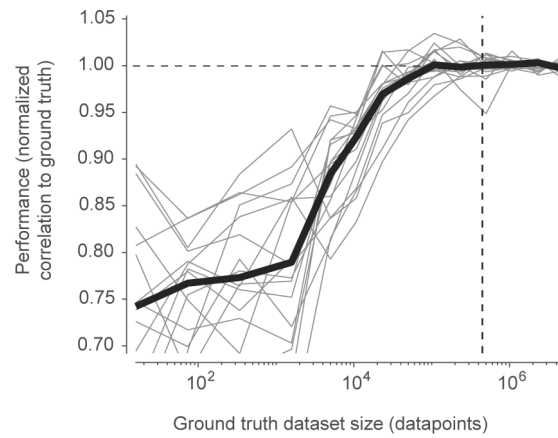


**Extended Data Fig. 4 | Generalization across neurons within a dataset.** The deep network was trained on all neurons of a specific dataset except one, and then tested with the remaining neuron. This analysis shows how the network is able to generalize to new neurons recorded under the same conditions, as a function of the standardized noise level  $\nu$  in  $\% \cdot \text{Hz}^{-1/2}$ . **a-d**, Performance of the predictions for 4 selected ground truth datasets in terms of correlation, error and bias as a function of the standardized 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 particularly low spike rates (black and green arrows) and particularly distinct calcium response kernel (red arrow, see main text for discussion). **e**, Correlation, error and biases as a distribution across neurons within each dataset (number of neurons for each dataset as indicated in Table 1). For box plots, the median is indicated by the central line, 25th and 75th percentiles by the box, and maximum/minimum values excluding outliers (points) by the whiskers. All datasets were re-sampled at a frame rate of 7.5 Hz.

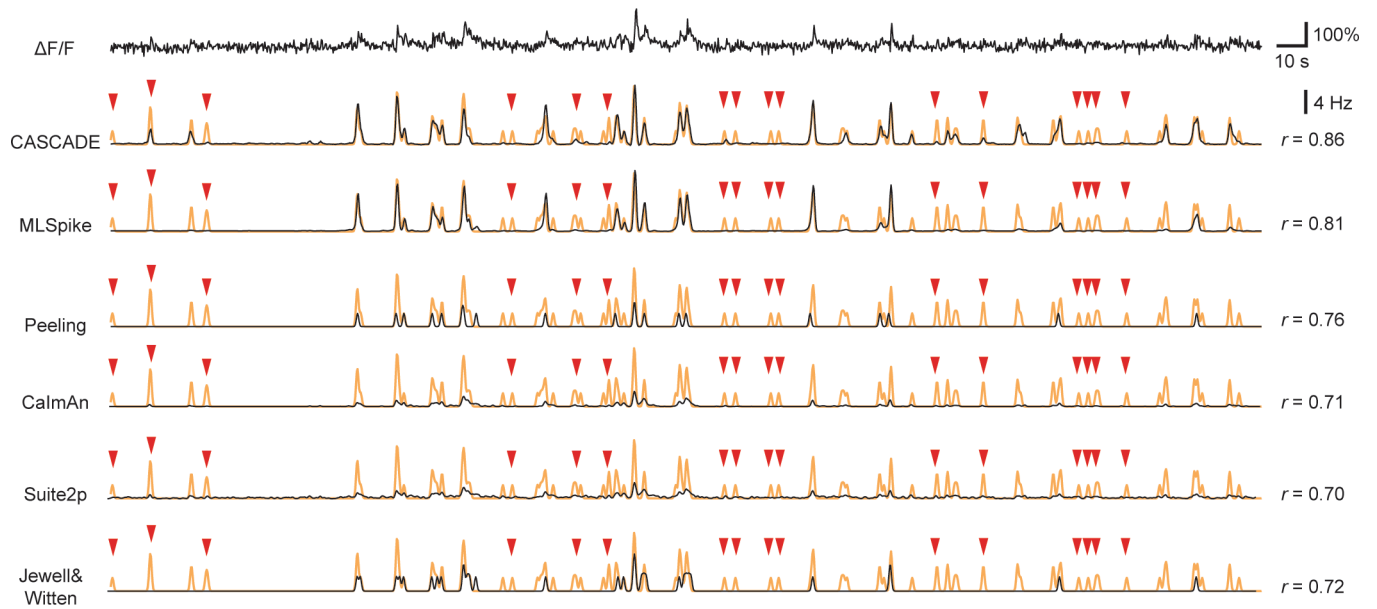


**Extended Data Fig. 5 | Typical artifacts in ground truth recordings.** Calcium trace ( $\Delta F/F$ ), true action potentials (APs), inferred spiking activity (SR) and true ground truth spiking activity (GT). **a**, The baseline of this recording is unstable, exhibiting irregular bumps (arrowheads). 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. Green arrowheads indicate movement artifacts that are not associated with high spiking activity (correct rejections of artifacts), while black arrowheads indicate movement artifacts that are not recognized as artifacts by the network (false positives). The zoom-in on the right shows an example where a movement artifact is associated with a negligible spike rate (correct rejection). **b**, Fluorescence transients without corresponding action potentials are clearly visible (red 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 (correct rejections). **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 anaesthetized animal underlying the movement artifact.

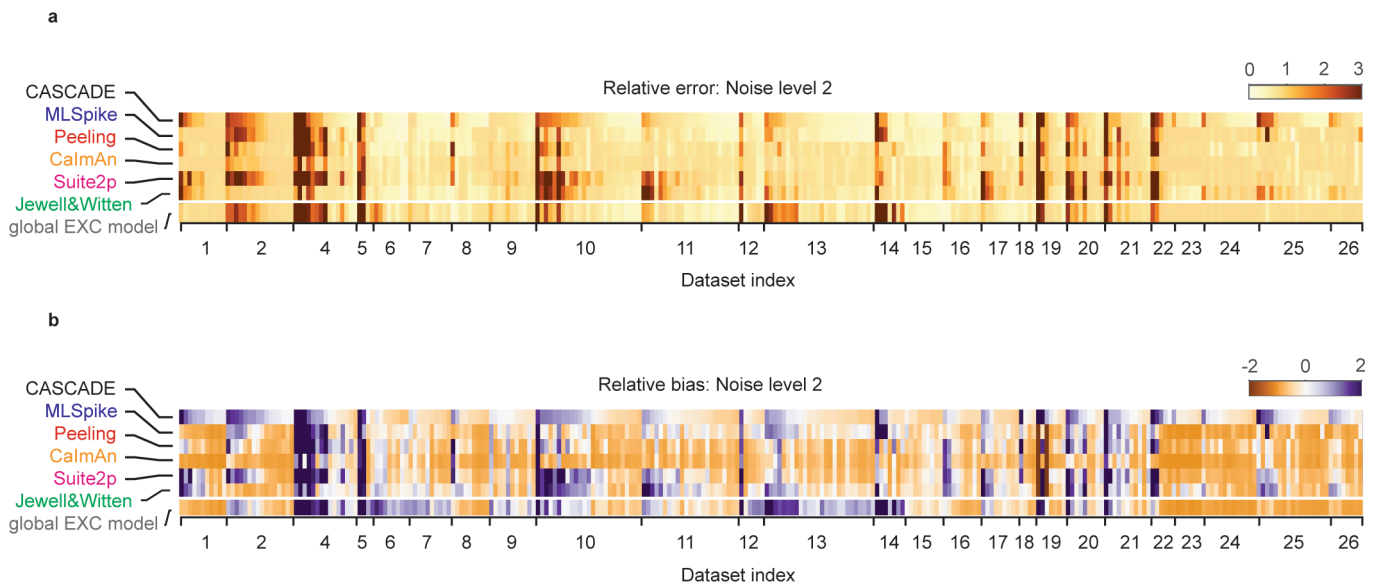




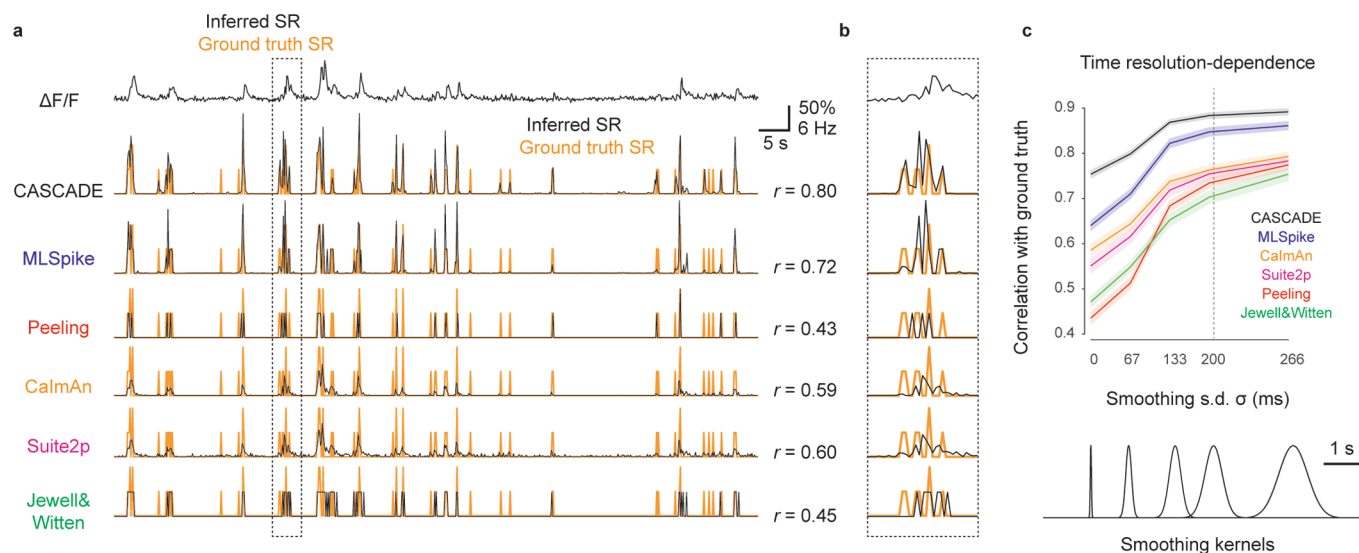
**Extended Data Fig. 6 | Improvement of performance with ground truth dataset size.** The global EXC model (see Fig. 3) was trained as before, but using only a subset of the ground truth data points (x-axis). The performance (correlation) across each dataset was normalized to the performance with 5 million data points (horizontal dashed line). The performance approaches an asymptote at approximately 100,000 data points. A typical single ground truth dataset contains ca. 400,000 data points (median across all datasets; vertical dashed line). This result also indicates that a diverse but smaller training dataset sampled from all ground truth datasets results in better generalization than a larger training dataset from a single ground truth dataset.



**Extended Data Fig. 7 | Comparison with model-based algorithms, extension of Fig. 4a.** Example predictions from the deep-learning based method (CASCADE) and five model-based algorithms (MLSpoke, CalmAn, Peeling, Suite2p, Jewell&Witten) of a  $\Delta F/F$  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. 4a, the calcium recording here is rather noisy due to the insensitivity of GCaMP to single action potentials in this neuron.

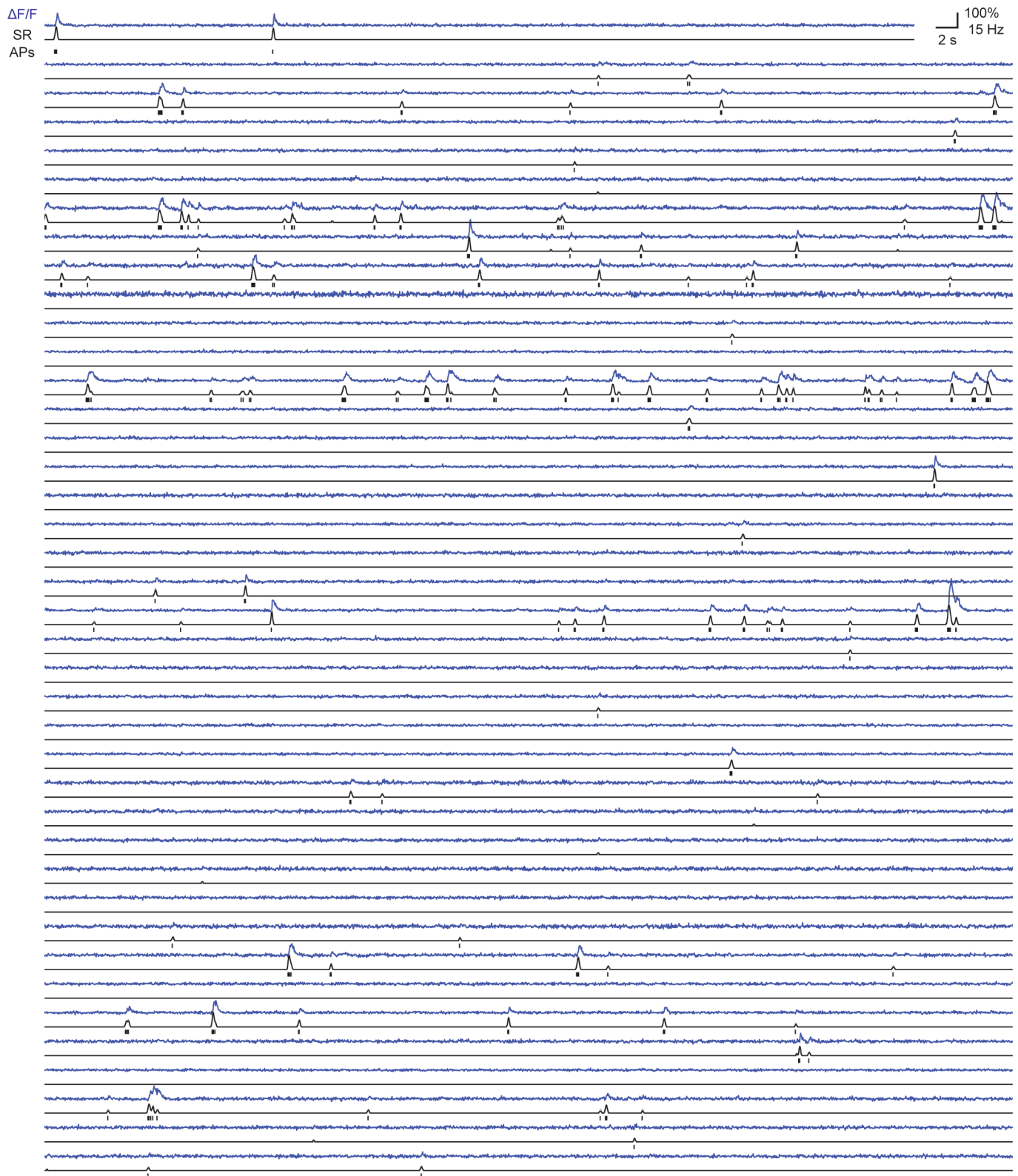


**Extended Data Fig. 8 | Comparison of CASCADE with model-based algorithms, extension of Fig. 4b.** Comparison of the six algorithms when optimized for a single dataset, showing relative error and relative bias for all neurons, grouped by ground truth dataset.



**Extended Data Fig. 9 | Performance dependence on temporal precision of predictions.** All algorithms were optimized via the mean squared error to infer spike rates at a specific temporal precision defined by the smoothing of the ground truth (default: Gaussian smoothing with kernel of  $\sigma = 200$  ms). For all model-based algorithms, the inferred spike traces were shifted in time to optimize the mean squared error. **a**, Predictions from an example  $\Delta F/F$  trace (top; dataset #09). Ground truth spike rates are shown in orange, inferred spike rates as black overlay. Correlation values are indicated at the right. The scale bars for  $\Delta F/F$  and time are the same as in Fig. 4a. **b**, Highlighted excerpt from (a). Due to the high temporal precisions of the inferred spike rates, small time shifts lead to low performance (clearly visible for the Peeling algorithm in this example). The CalmAn and Suite2p algorithms deconvolve less aggressively, therefore making less dramatic errors. CASCADE and MLSpike perform best for this example neuron, with CASCADE detecting more events than MLSpike. **c**, Overall performance (correlation) change with temporal precision of predictions (smoothing kernels shown below) on a subset of datasets (datasets #4, #6, #9, #11-14 and #18). As expected, correlation with ground truth decreased with higher temporal resolution of the desired temporal resolution. This decrease was especially prominent for algorithms that, by design, aim at the inference of precise (discrete) spike rates (Peeling, Jewell&Witten). The decrease was less pronounced for CASCADE compared to for example MLSpike. Shaded corridors indicate SEM across  $n = 8$  datasets. All recordings resampled at a noise level of 2 with a frame rate of 7.5 Hz.





**Extended Data Fig. 10 | Predictions of spiking probabilities and discrete spikes from the Allen Brain Institute Visual Coding dataset.** Predictions were produced with the global EXC model trained at 30 Hz. 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).

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

#### Data collection

For the collection of datasets 4-8, the Matlab-based open software Ephus (version described in Suter et al., 2010) was used for electrophysiology and Scanimage 3.8 (Pologruto et al., 2003) for calcium imaging. For collection of datasets 18-19, Clampex 10.2 software was used for electrophysiology and Helioscan (Langer et al., 2013) for calcium imaging. For dataset 1, AxoGraph software (version X) was used for electrophysiology and Scanimage (Pologruto et al., 2003) for calcium imaging. For the multiplane population imaging dataset in adult zebrafish (Figure 5), a customized version of Scanimage (Pologruto et al., 2003; Rupprecht et al., 2016; <https://github.com/PTRRupprecht/Instrument-Control/tree/master/ScanImage%20B>) was used. For dataset 24-27, WinWCP (John Dempster, University of Strathclyde) was used for electrophysiology and a custom software written in Labview for calcium imaging (Khan et al., 2018). For datasets 2, 22 and 23, Clampex (Molecular Devices) was used for electrophysiology and Scanimage (Pologruto et al., 2003) for calcium imaging.

#### Data analysis

The main data analyses including the implementation of the spike inference algorithm were performed in Python 3. The code for the spike inference algorithm together with extensive documentation is deposited on a Github repository (<https://github.com/HelmchenLabSoftware/Cascade>). Additional analyses and visualizations were performed in Matlab. Statistical tests were performed either in Matlab or in R.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The processed raw data of the ground truth datasets described in this study are available from Github (<https://github.com/HelmchenLabSoftware/Cascade>). Pretrained models of CASCADE are automatically retrieved by the code provided in the Github repository and are stored in the cloud ([www.switch.ch/drive](http://www.switch.ch/drive), via University of Zurich). The raw unprocessed data that support the findings of this study are available from the corresponding author upon request.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were not predetermined for this study. For each dataset, ground truth recordings from 10-20 neurons of a minimum duration would be desirable to provide a clear measurement of variability across a dataset. However, fewer neurons were recorded from some datasets due to technical difficulties, and the resulting datasets were also useful for our analyses. In addition, this study explicitly analyzed the effect of limited sample sizes for each of the analyzed datasets (see Extended Data Figure 6).
Data exclusions	Exclusion of low-quality experimental ground truth data was an essential part of this study and is described in detail in the Methods section "Quality control". Simultaneous calcium imaging and juxtacellular recordings are technically challenging experiments that can often result in suboptimal recordings. Suboptimal recordings with, e.g., strong movement artifacts during imaging or contamination of the juxtacellular recordings by spikes from adjacent neurons would result in erroneous ground truth and therefore requires quality control including the exclusion of such low-quality recordings.
Replication	All results and analyses based on the spike inference algorithm (CASCADE) were replicated on the same computer and on other computers on various operations systems (Linux, Windows, Mac). Ground truth experiments were not replicated beyond the recordings shown in Table 1. Multiplane population imaging experiments in adult zebrafish (Fig. 5) were replicated with similar results in >20 fish.
Randomization	Animals were not allocated into different groups in this study.
Blinding	Animals were not allocated into different groups in this study.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	For experiments with zebrafish, wildtype or genetically modified (neuroD:GCaMP6f, Rupprecht et al., 2016) adult zebrafish (5-10 months old) of both sexes were used. For experiments for dataset 1, C57BL6/J male mice (postnatal days 28–61) were used. For experiments for datasets 17 and 18, 8-12 week-old C57BL/6J male mice and genetically modified (Grik4-cre-G32-4Stl) male mice were used. Male crosses between PV-Cre and Rosa-CAG-LSL-tdTomato were used for dataset 24 and 27, VIP-Cre male mice for dataset 25 and SOM-Cre female mice for dataset 26 (all mice postnatal days 36-56, except for dataset 27, where mice were >8 weeks old). Male GIN mice (FVBTg(GadGFP)45704Sw/J) were used for dataset 22, male crosses between a PV-Cre mouse [(B6;129P2-Pvalbtm1(cre)Arbr/J) and a loxP-flanked tdTomato reporter mouse (B6.Cg-Gt(ROSA)26Sortm14(CAG-tdTomato)Hze/J) were used for dataset 23, and male crosses between a CaMKII $\alpha$ -Cre mouse (B6.Cg-Tg(Camk2a-cre)T29-1Stl/J) and a loxP-flanked tdTomato reporter mouse were used for dataset 2.
Wild animals	No wild animals were used in this study.
Field-collected samples	No samples were collected from the field for this study.
Ethics oversight	All experimental procedures for experiments with mice (datasets 18, 19) 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. The care of animals and experimental procedures for dataset 1 were carried out in accordance with national and institutional guidelines, and all experimental protocols were approved by the Animal Experimental Committee of the University of Tokyo. All zebrafish experiments (datasets 4-8) and the experiments for datasets 24-27 were approved by the Veterinary Department of the Canton Basel-Stadt (Switzerland). All experimental procedures for datasets 2, 22 and 23 were performed in accordance with NIH guidelines and approved by the Animal Care and Use Committee at University of California, Berkeley.

Note that full information on the approval of the study protocol must also be provided in the manuscript.