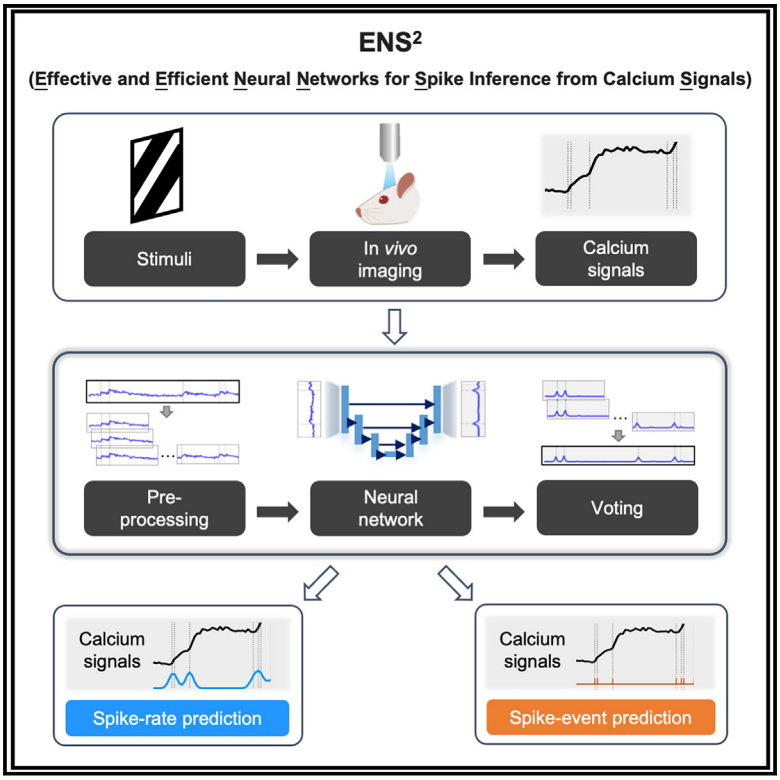
Cell Reports
Methods

Article

Effective and efficient neural networks for spike inference from *in vivo* calcium imaging

Graphical abstract



Highlights

- ENS^2 uses a U-Net architecture with MSE loss for effective spike inference
- ENS^2 is calibration free and handles calcium signals with varying signal-to-noise ratios
- ENS^2 enables characterization of orientation selectivity of V1 neurons
- The calcium signal transient amplitude is a key determinant of inference performance

Authors

Zhanhong Zhou, Hei Matthew Yip, Katya Tsimring, Mriganka Sur, Jacque Pak Kan Ip, Chung Tin

Correspondence

jacqueip@cuhk.edu.hk (J.P.K.I.), chungtin@cityu.edu.hk (C.T.)

In brief

Zhou et al. develop a spike inference system from calcium signals based on a U-Net deep neural network. It is calibration free and computationally efficient. It performs consistently well with calcium signals of variable signal quality with a single model and facilitates analyses of orientation selectivity in V1 neurons.

Zhou et al., 2023, Cell Reports Methods *3*, 100462
May 22, 2023
https://doi.org/10.1016/j.crmeth.2023.100462

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Effective and efficient neural networks for spike inference from *in vivo* calcium imaging

Zhanhong Zhou,[1] Hei Matthew Yip,[2] Katya Tsimring,[3] Mriganka Sur,[3] Jacque Pak Kan Ip,[2,*] and Chung Tin[1,4,*]
[1]Department of Biomedical Engineering, City University of Hong Kong, Hong Kong SAR, China
[2]School of Biomedical Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China
[3]Department of Brain and Cognitive Sciences, Picower Institute for Learning and Memory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[4]Lead contact
*Correspondence: jacqueip@cuhk.edu.hk (J.P.K.I.), chungtin@cityu.edu.hk (C.T.)
https://doi.org/10.1016/j.crmeth.2023.100462

MOTIVATION Calcium imaging is a powerful tool for monitoring the activities of multiple neurons and understanding brain function, but its generally low signal-to-noise ratio (SNR) and slow dynamics limit its precision and temporal resolution compared with the traditional electrophysiological recording. To overcome this limitation, reliable models for spike inference from calcium signals will be beneficial. Here, we aim to develop efficient and calibration-free spike inference models that generalize well to a broad range of calcium data, including various calcium indicators, SNRs, brain regions, and so on.

SUMMARY

Calcium imaging provides advantages in monitoring large populations of neuronal activities simultaneously. However, it lacks the signal quality provided by neural spike recording in traditional electrophysiology. To address this issue, we developed a supervised data-driven approach to extract spike information from calcium signals. We propose the ENS^2 (effective and efficient neural networks for spike inference from calcium signals) system for spike-rate and spike-event predictions using $\Delta F/F_0$ calcium inputs based on a U-Net deep neural network. When testing on a large, ground-truth public database, it consistently outperformed state-of-the-art algorithms in both spike-rate and spike-event predictions with reduced computational load. We further demonstrated that ENS^2 can be applied to analyses of orientation selectivity in primary visual cortex neurons. We conclude that it would be a versatile inference system that may benefit diverse neuroscience studies.

INTRODUCTION

One key to understanding the complex functions of the brain is to simultaneously measure the activity of neurons across different layers and brain areas. Electrophysiological recordings, such as patch-clamp[1,2] and multielectrode extracellular recording,[3] have long been the major method to record neuronal spiking events. These recordings are typically of high temporal resolution and with high signal-to-noise ratio (SNR). However, it is technically challenging with these methods to acquire recordings from a large number of neurons stably *in vivo*.[4]

In recent decades, the optical-based two-photon calcium imaging technique has increasingly been used for *in vivo* neuroscience research.[5–12] This imaging technique enables simultaneous monitoring of activities of thousands of neurons over a considerable period of time. Moreover, as more effective fluorescent calcium indicators[13–17] and imaging devices[18–20] have become available, it is now possible to localize and extract the individual activities of a large number of neurons in various subcellular structures.[21]

Nevertheless, calcium imaging is only an indirect measurement of neuronal activities. In brief, the concentration of intracellular calcium evoked by neuronal firings undergoes nonlinear changes. These fluctuations in calcium are again nonlinearly reflected by calcium indicators, whose fluorescent intensities could be imaged. Afterward, the locations of individual neurons or compartments (region of interest [ROI]) are identified on images, and the time-varying fluctuations of fluorescence signals in the ROIs are extracted as a surrogate of neuronal activities. Another limitation of calcium imaging is that the signals can commonly have a low SNR,[21] especially for those recorded in deep brain regions *in vivo* or at low signal conditions. Furthermore, the indicators' slow temporal dynamics up to hundreds of milliseconds[14,22] would result in low-pass-filtered activities. These indicators come in different types, typically synthetic dyes or genetically encoded calcium indicators (GECIs), and their different dynamics further complicate the task to convert the calcium signal into neuronal signals.

Previous work has shown that spike inference plays a crucial role in interpreting calcium data and dissecting neural circuits.[5–8] In the past decades, researchers have developed various algorithms to recover multiunit neuronal spikes. These algorithms can be generally divided into two major categories: model-based systems[23–34] and data-driven systems.[35–39] In model-based systems, physiologically constrained models were typically built, considering that the calcium signal concentrates with neuronal firings and decays exponentially afterward. With these models, calcium traces could be simulated through estimated spike trains and additive noises. They include systems based on template matching[24,25,27] (e.g., peeling[25]), deconvolution[23,26,30,32–34] (e.g., OASIS[30]), and Bayes' theorem[26,28,29,31] (e.g., MLspike[29]). For example, as one of the state-of-the-art algorithms, MLspike was proposed using a physiologically constrained model and optimized by maximum *a posteriori* (MAP) estimate to infer the most likely spike trains from noisy calcium signals. However, these model-based methods typically require tuning of model parameters for each new recording, either manually or to be estimated by auxiliary algorithms. Moreover, when likelihood optimization is involved, they may become rather computationally expensive to use. On the other hand, data-driven systems based on supervised learning also emerged with promising performances. A supervised deep learning algorithm called CASCADE has been reported recently, which delivered high spike inference performance when training data with matched noise levels as the testing neurons were selected for training.[38] Previous data-driven models have faced challenges in validating their generalization ability because of the limited high-quality paired data for training.[35,36,39] An extensive public database of paired data (simultaneously recorded calcium fluorescence signals and electrophysiology ground truths) has been compiled alongside the development of CASCADE. It has facilitated such data-driven approaches for better generalization of the models, although some re-training is still necessary for noise matching.[38] Some other works also use feature extraction and thresholding[40,41] (e.g., GDspike[41]) to tackle the problem of spike inference.

For inferring unpaired calcium signals from *in vivo* imaging, a calibration-free inference system that could generalize on unseen recordings with high performance is desirable. In fact, neural networks have shown satisfactory performance in processing bio-signals with severe inter-record variability, including electrocardiogram[42–44] and electromyography.[45] Provided with a sufficient amount of paired data, the generalization ability of neural networks makes it a promising approach for inferring spikes from calcium signals. In this work, we performed thorough research on the impact of each component in the neural network-based system on the spike inference tasks, based on a large ground-truth public database (Table S1). The optimal configurations of network architectures and cost functions were investigated. We conducted additional simulations to address factors in the calcium data that could benefit the performance of deep-learning-based models for spike inference. These analyses provided useful insights on how to prepare (e.g., record, process, and select) calcium data that will favor future algorithm development, which could help us understand the complex process in the brain. Here, with these research and insights, we developed the ENS^2 (effective and efficient neural networks for spike inference from calcium signals) system (Figure 1) with state-of-the-art performance and generalization ability but with lower computational complexity. To further demonstrate the validity of the ENS^2 system, we deployed the ENS^2 system on a set of calcium imaging data from the primary visual cortex (V1) and showed how the spike inference can be applied to the analyses of the experimental data.

RESULTS

Design of neural network-based spike inference system

Network architectures

We tested three different architectures of neural networks to evaluate their effectiveness in the spike inference task. They are U-Net,[46] Le-Net,[47] and FC-Net (fully connected network), respectively. These existing models were modified for 1D calcium signal inputs. The network architectures are summarized in Table S2.

The U-Net used in this study contains three contracting blocks and expanding blocks. On one hand, the input information from contracting blocks passes through the bottleneck block to the expanding blocks. On the other hand, skip connections from contracting blocks to the corresponding expanding blocks allow direct and localized information flows.[46] Within each contracting/expanding block and the bottleneck block, two convolution layers with 3-sized kernels are deployed. Instead of batch normalization, we used instance normalization[48] for regularization, because calcium signals with various dynamics may co-exist in a same batch of data. We observed that this regularization helped in model convergence. A schematic of the proposed U-Net architecture is shown in Figure 1C. The Le-Net consists of three convolution layers with kernel sizes of 3, 3, and 10, respectively. Average pooling layers with 2-sized kernels are applied between the three convolution layers. A dropout layer[49] is included before the output layer for regularization. A typical fully connected network with four hidden layers and two dropout layers is adopted as FC-Net. All three networks are designed to take 96-sized calcium signal inputs and output 96-sized spike-rate vectors in a sequence-to-sequence translation manner. Given that the input data are segmented with steps of one data point, the spike rate at each time point is indeed predicted for up to 96 times independently from its adjacent segments (zero padding was performed at both the beginning and the end of the recordings before segmentation). We then average these 96 predictions to provide the final spike-rate output of each time point (Figure 1A). These 96 segments can provide long- and short-range information about a specific time bin. Hence we can achieve more robust inference results with such expanded "receptive field" after averaging. Spike-event output can then be estimated from this final spike-rate sequence as introduced below.

We kept all three networks to have similar numbers (under 150k) of trainable parameters for comparison. They were all randomly initiated to have zero means and standard deviations of 0.02. Leaky ReLUs (rectified linear units) with slopes of 0.2 are used as activation functions for all layers except for the output layers, where ReLU is used for non-negative spike-rate prediction.

CellPress
OPEN ACCESS

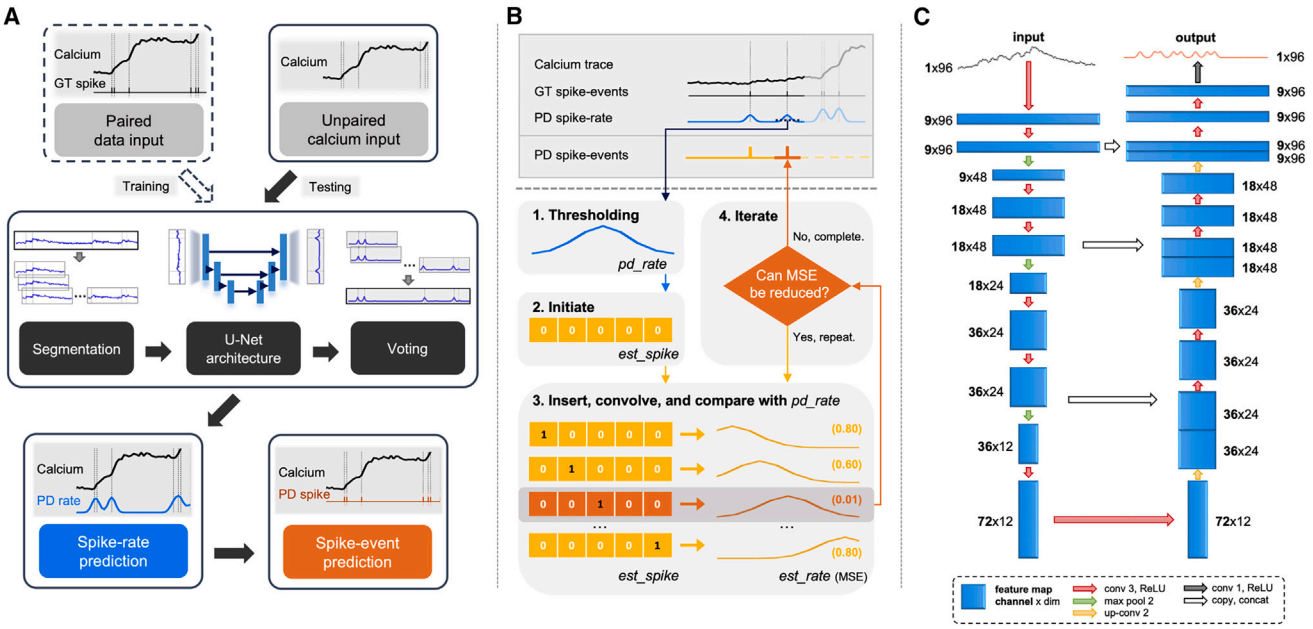


Figure 1. Overall workflow of the proposed ENS^2 system

(A and B) The ENS^2 system contains a neural network to infer spike rate from calcium inputs (A) and an unsupervised greedy algorithm for estimating spike events from spike-rate predictions (B). The neural network is trained with calcium trace inputs paired with ground-truth (GT) spike events, whereas it could test on calcium traces alone after training for obtaining predicted (PD) spike rates. For a given calcium recording, our ENS^2 system will first predict the corresponding spike rate with the procedures in (A).

(C) 1D calcium trace inputs are segmented to have 96 data points and fed to the U-Net-based model. It outputs the corresponding spike-rate prediction of the same length in a sequence-to-sequence manner, which is gathered through an averaging strategy. Afterward, spike events are estimated by the four-step algorithm in (B). In brief, valid fragments of spike-rate prediction are extracted by thresholding. Then, estimated spike-events sequences (*est_spike*) are formed by tentatively inserting spike event in each time bin. The spike-events sequences (*est_spike*) are convolved with a smoothing window to approximate the spike-rate segment (*est_rate*). The sequence with minimum MSE against the prediction (*pd_rate*) is regarded as the final spike-events prediction. Details are explained in the results section. Detailed hyper-parameters are summarized in Table S2.

(conv, convolution; ReLU, rectified linear unit; max pool, maximum pooling; up-conv, transposed convolution; concat: concatenation).

Loss functions and optimization

For each type of network, we optimized them with three different loss functions, respectively, for comparison. First, mean square error (MSE) loss is used, which is one of the most commonly used loss functions applicable to a wide variety of machine learning tasks. The models are expected to minimize the MSE between predicted spike rates and ground-truth spike rates, penalizing the prediction both in time and amplitudes. The loss function is as follows:

$$\mathcal{L}_{MSE} = \frac{1}{n}\sqrt[2]{(GT - PD)^2}, \qquad \text{(Equation 1)}$$

where n is the number of segments in a batch, and GT and PD stand for ground truth and prediction, respectively. In addition, we also used Pearson correlation coefficient (Corr) and van Rossum distance (vRD)[50] as loss functions:

$$\mathcal{L}_{Corr} = \frac{E[(GT - \mu_{GT})(PD - \mu_{PD})]}{\sigma_{GT}\sigma_{PD}} \qquad \text{(Equation 2)}$$

$$\mathcal{L}_{vRD} = \sqrt{\frac{1}{\tau}\frac{\int [GT(t) - PD(t)]^2 dt}{\int [GT(t)]^2 dt}}, \qquad \text{(Equation 3)}$$

where μ and σ are the mean and standard deviation, respectively, and τ is a constant normalizing factor (see the "evaluation of spike inference algorithms" section in STAR Methods). Note that Corr and vRD are also used as the metrics for evaluating the performance of the models (see STAR Methods). As such, we would expect that models that use $\mathcal{L}_{Corr}$ (or $\mathcal{L}_{vRD}$) should have the optimal performance when measured with Corr (or vRD). A major difference between $\mathcal{L}_{MSE}$ and $\mathcal{L}_{vRD}$ is that $\mathcal{L}_{vRD}$ normalizes each batch of samples by the total numbers of GT firing events. As such, the optimization through $\mathcal{L}_{vRD}$ is less dependent on the firing rates of the training data but is slightly more computationally expensive to use than $\mathcal{L}_{MSE}$. We will compare their resultant inference performance in further detail.

The Adam optimizer[51] with a default learning rate of $1e^{-3}$ is used for all models. Each model is allowed to update for a maximum of 5,000 iterations. In each iteration, a batch of 1,024 paired segments is drawn randomly and fed to the model for training. The training losses are noted, and early stopping is introduced when the losses do not improve in the past 500 interactions (*patience* = 500). Under these criteria, we observed that most models completed the trainings within 3,000 iterations. The resultant models are then ready for prediction. Other details of

hyper-parameters and operational environment are summarized in Table S3.

Estimation of spike events from spike-rate predictions

To reliably convert the spike-rates output by the neural networks to spike-event predictions, we propose an unsupervised greedy algorithm that is simple and straightforward (Figure 1B). The workflow is briefly introduced here.

Step 1: Fragments of spike-rate predictions (*pd_rates*) with non-zero spike rate are identified by thresholding the spike-rate sequence output with an epsilon value. We do not use zero threshold to avoid including any fragment with overly low peak amplitude (i.e., those showing extremely small spiking probabilities or background noise), where no spike should be estimated.

Step 2: For each *pd_rate* of length *L* (in terms of number of data points), we initialize a zero-filled vector (*est_spike*) with *L* bins.

Step 3: One spike is assigned to any one bin in *est_spike* at one time. Then the *est_spike* vector is convolved into a spike-rate vector (*est_rate*) with the smoothing windows (as in the pre-processing step described in Methods). The corresponding MSE between the resultant *est_rate* and *pd_rate* is calculated. This step is implemented in parallel to all *L* bins to determine the most suitable bin (i.e., with the smallest MSE) for assigning the spike.

We then repeat step 3 to assign another spike each time to the most suitable bin in a greedy manner, until the MSE would no longer be reduced by adding a spike to any location in *est_spike*. Then the updated *est_spike* is regarded as the final estimation of spike events for the concerned *pd_rate* fragments. The time stamp of a spike is defined as the center time of the corresponding bin within *est_spike*. If multiple spikes are predicted in the same bin, the same time stamp is repeated accordingly.

For a spike-rate sequence output with N fragments of *pd_rates*, this algorithm executes in $O(N \times L \times k)$ time, where k is the maximum number of spikes in any one bin. In practice, considering the typically slow dynamics of calcium signals and relatively low firing rates of neurons imaged, this estimation method operates in linear time proportional to the duration of recordings. We have validated our system with this spike-events estimation algorithm against several existing studies, including MLspike,[29] OASIS,[32] and CASCADE[38] with its Monte Carlo importance sampling-based spike-events estimation algorithm.

U-Net and MSE loss achieve the best overall performance in spike inferring tasks

As described above, our benchmark involved configurations from three types of neural network model and three loss functions, resulting in a total of nine configurations of models. Our simulations followed the leave-one-dataset-out protocol. For example, when benchmarking on excitatory neurons, each time the model was first trained on 19 datasets and tested on the remaining one. This was repeated 20 times such that all 20 datasets (1–20) with excitatory neurons were tested respectively. The procedure is similar for the six datasets (21–26) with inhibitory neurons. We then recorded the neuron-wise performance in all 26 datasets. The results are presented for each neuron from the testing datasets (Figures 2A–2D). All datasets were re-sampled to a frame rate of 60 Hz for benchmark (see STAR Methods).

First, we compared the three architectures of neural networks (U-Net, Le-Net, and FC-Net, see above). As shown in Figures 2A–2D, U-Net delivered the best overall performance against either Le-Net or FC-Net ($p < 0.0001$ for all cases, when MSE or vRD loss function was used). Next, we assessed how different loss functions would affect the performance of our models. Here, MSE, vRD, and Corr were used as the loss functions, respectively. The vRD and Corr were used as loss function to test whether they would favor spike-rate prediction because they are also evaluated by vRD and Corr. Nevertheless, our results show that MSE loss appeared to be the better choice in general. When measured in Corr (Figure 2A), using MSE loss showed similar performance with using Corr loss ($p = 0.52$) and obtained higher Corr compared with vRD loss ($p = 0.0006$). When measured in vRD (Figure 2B), using MSE loss also showed similar performance with using vRD loss ($p > 0.99$) and obtained lower vRD compared with when using Corr loss ($p < 0.0001$). For spike-event predictions, using MSE loss again gained advantages over using vRD loss in error rate (ER) measurement ($p = 0.0032$; Figure 2C), and comparable performance in Victor-Purpura distance (VPD) measurement ($p \geq 0.99$; Figure 2D). Note that when Corr loss was used, the performance of all three models was significantly compromised, except when measured in Corr. The degradation in performance was most prominent in spike event predictions (measured in ER and VPD; Figures 2C and 2D). We believe that the major reason is that the Corr is a scale-free measurement, and Corr loss function fails to differentiate predictions of different amplitudes (but only different temporal patterns). As shown in Figure 2G, the prediction obtained with models using Corr loss tended to have much lower spike rate (in amplitude) as compared with other configurations (Figures 2E–2G). As a consequence, spike event could not be reliably estimated from the predicted spike rate due to low SNR. Putting the above results together, we took U-Net and MSE loss function in our proposed ENS^2 as it achieved the best overall performance in the benchmark.

To show that the difference in performance resulted from the model configurations rather than specific hyper-parameter settings, we repeated the simulation with U-Net and MSE loss with various hyper-parameters (Figure S1). The filled bars represent the default hyper-parameter combination used in this study as described above. Regardless, we showed that they all had little effect on the final performance. The Corr approached 70%, and vRD remained less than 3 in all cases. The VPD and ER were around 0.6 and 50%, respectively.

We also proved that our models were trained adequately with our early-stopping criteria (see above). Figure S1G illustrates the MSE training losses for all 20 datasets with excitatory neurons (red) and all 6 datasets with inhibitory neurons (blue). The losses decreased with more iterations generally and stabilized sufficiently as the training stops. Moreover, Figure S1 demonstrates that the patience of iterations (see above) before early stopping had little influence on performance. Together, we proved that our

CellPress
OPEN ACCESS

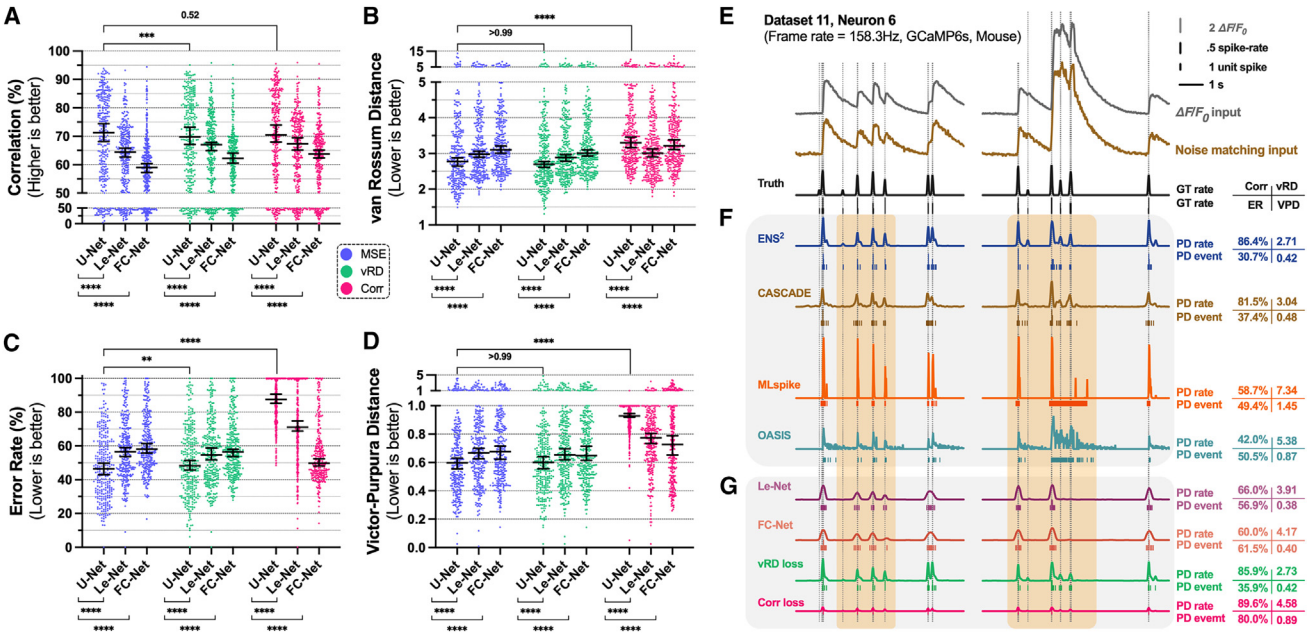


Figure 2. Performance comparison among various model architectures and loss functions

(A–D) Performances of various system configurations measured in (A) correlation, (B) van Rossum distance, (C) error rate, and (D) Victor-Purpura distance, respectively. Generally, the configuration of U-Net trained with MSE loss provides the best overall performance for both spike-rate prediction and spike-event prediction.

(E) Examples of different (pre-processed) calcium inputs (colored), paired with ground-truth (GT) spike rate and spike events (black).

(F) Examples of spike rates and spike events predicted (PD) by our proposed ENS^2 method and state-of-the-art methods.

(G) Examples of spike rates and spike events predicted by our proposed method with different configurations.

All error bars in this work represent medians with 95% confidence intervals. Asterisks indicate significant difference using Friedman's test with Dunn's multiple comparisons between the indicated configurations (see STAR Methods, n = 286). *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001).

Colored dots in (A)–(D) represent measured performance of each neuron. Metrics shown in (F) and (G) measure the performance on the corresponding neuron. Orange shaded areas represent regions of interest where discrepancies in predictions are observed among different methods. These conventions are consistent for the remaining figures.

models have been trained and regularized sufficiently by iterating over only thousands of batches of data to avoid over-fitting.

Comparison with state-of-the-art models

Based on our investigations above, we selected the configuration of U-Net and MSE loss as our proposed ENS^2 system, which takes original $\Delta F/F_0$ signals as inputs. We took this further to compare it with three representative state-of-the-art studies, including CASCADE,[38] MLspike,[29] and OASIS.[32] We selected CASCADE and MLspike as they are among the top-performing systems within the two major categories: data-driven systems and model-based systems, respectively. Moreover, both of them have already shown surpassing performance over previous methods using various datasets and evaluation metrics in their studies. On the other hand, OASIS is one of the most representative methods based on deconvolution and has been implemented in Suite2P[52] and CaImAn[53] for wide experimental usages. Results are summarized in Figure 3.

We first benchmarked their performance across all 26 datasets (see STAR Methods). Figures 3A–3D shows that the data-driven systems (i.e., our proposed ENS^2 and CASCADE) generally performed better than the model-based system (e.g., p < 0.0001 for all cases of ENS^2 vs. MLspike/OASIS) in both spike-rate (Corr and vRD) and spike-event (VPD and ER) predictions. For example, our ENS^2 showed around 34%/36% higher in Corr and 18%/15% lower in ER than MLspike/OASIS. When compared with CASCADE, our systems also showed better performance for both spike-rate prediction and spike-event prediction (p < 0.0001 for all cases; Figures 3A–3D, S2C, and S2D). In particular, our ENS^2 showed around 5% higher in Corr and 4% lower in ER than CASCADE.

We took a deeper look into these results by considering the excitatory neurons and inhibitory neurons separately (Figure S3). Generally speaking, all four systems performed worse in inhibitory neurons than in excitatory neurons (see discussion). Notwithstanding, the ENS^2 system presented better results in inhibitory neurons than the other models (Corr: p = 0.013 for ENS^2 vs. CASCADE, p < 0.0001 for the rest; vRD: p = 0.0093/0.0018 for ENS^2 vs. CASCADE/OASIS, p < 0.0001 for the rest; ER: p < 0.0001 for ENS^2 vs. MLspike; VPD: p < 0.0001 for ENS^2 vs. MLspike/OASIS). For example, our ENS^2 showed around 11%/54%/27% higher in Corr and 6%/40%/2% lower in ER than CASCADE/MLspike/OASIS for inhibitory neurons. These results proved that our ENS^2 system yielded better performance consistently for both excitatory and inhibitory neurons.

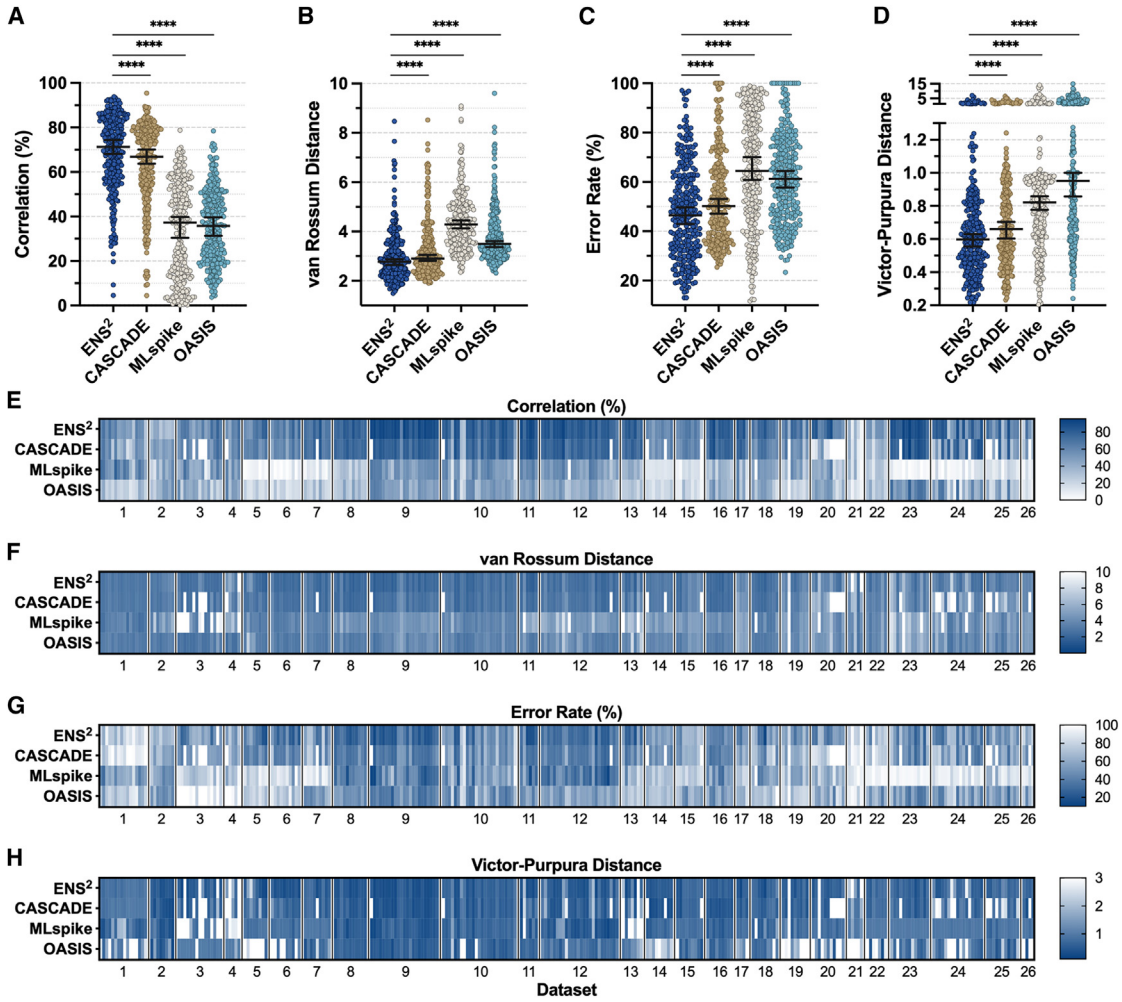


Figure 3. Performance comparison of our proposed ENS^2 system against state-of-the-art models

(A–D) Neuron-wise performance is measured in (A) correlation, (B) van Rossum distance, (C) error rate, and (D) Victor-Purpura distance, respectively.

(E–H) Inference performance on each neuron in all datasets is summarized in (E)–(H) heatmaps.

Colored circles in (A)–(D) denote the performance for each individual neuron. Error bars represent medians with 95% confidence intervals. Asterisks indicate significant difference using Friedman's test with Dunn's multiple comparisons between the indicated systems (****$p < 0.0001$).

We also investigated how the performance of each algorithm varied for each single neuron (or dataset) (Figures 3E–3H). The comparison showed that neurons that were well/badly predicted by one algorithm were usually predicted (relatively) well/badly by the other algorithms as well (e.g., better on datasets 9–13 but worse on datasets 14–15). This suggests that the performance of inference for certain neuron (or dataset), regardless of algorithm, would depend significantly on their own properties. We explored the underlying factors to such phenomenonin further details in the section "Factors affecting inference performance" below.

We selected several specific segments of recording from five different neurons (Figures 2E, 2F, and 4A–4D) for further comparison among the four algorithms. In Figure 2E, the recording with GCaMP6s indicator has a relatively high frame rate of ~158 Hz, and the original calcium trace (in $\Delta F/F_0$) has large amplitudes with only small noise. Figure 2F shows that CASCADE tended to output broader spike-rate prediction and thus longer sequences of spike events than ENS^2. This may be because the "noise-matching input" used by CASCADE could contain more noise than the actual input because of rounding of noise level to integer values in the algorithm. As such, the model perceived more noise than actually existing in the inputs. In contrast, both MLspike and OASIS significantly over-estimated with long sequences of spike events. Overall, our system (ENS^2) showed better predictions than these three methods for all evaluation metrics (values on the right in Figure 2F). Figure 4A shows another sample recording of high SNR but with GCaMP6f indicator. In this case, our ENS^2 system again recovered the spike-rate and spike-event patterns the most properly, whereas the other three systems tended to under-estimate and missed several spikes. In contrast, when the frame rate of the recording was decreased to 30 Hz with lower SNR (Figure 4B), the inference task became more challenging. We observed slight over-estimation from the ENS^2 system and several missed spikes from the

CellPress
OPEN ACCESS

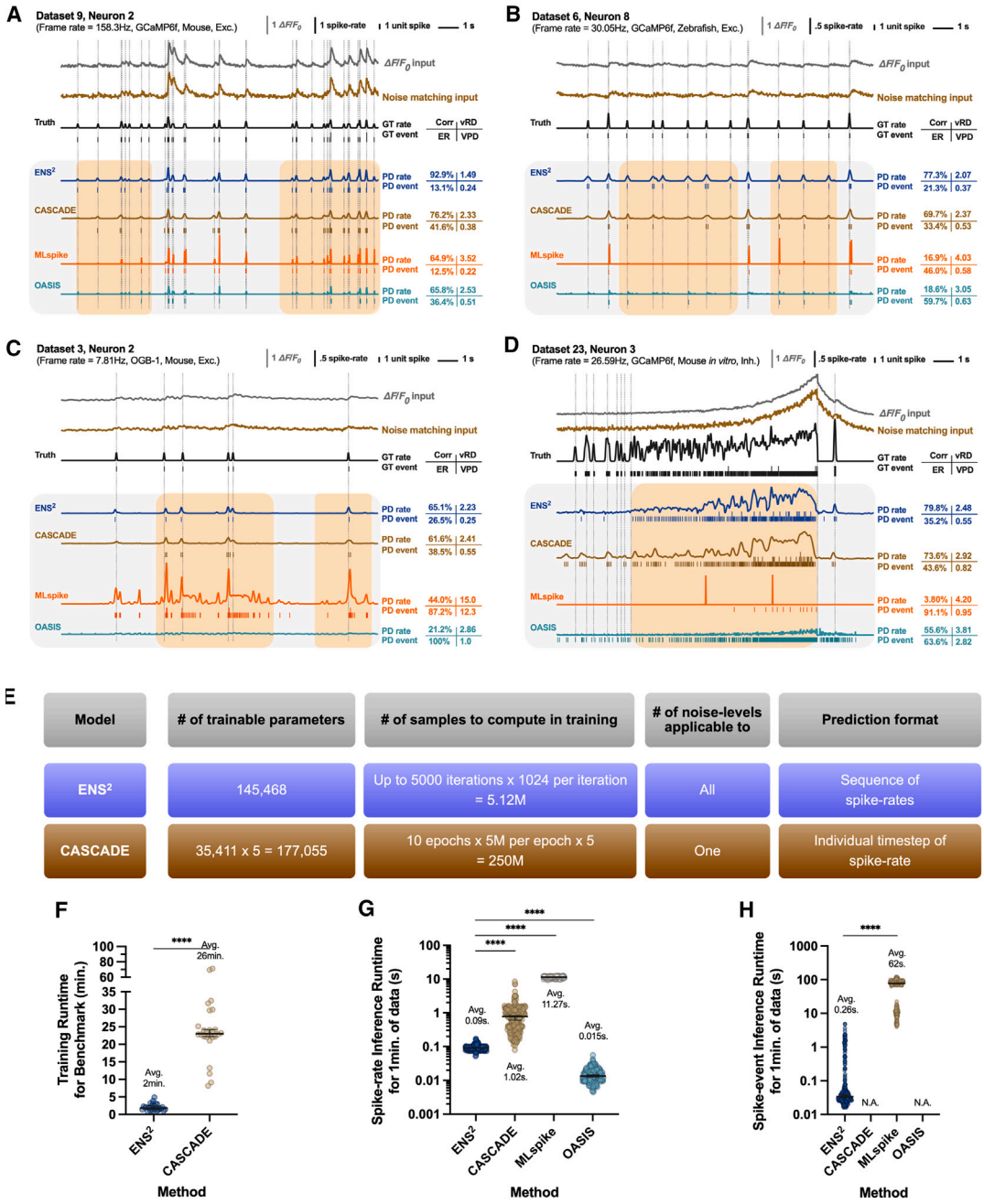


Figure 4. Examples of spike-rates and spike-events prediction, computational complexity, and run time of our proposed ENS^2 and state-of-the-art models

(A–D) Spike-rates and spike-events prediction and the corresponding performance measured from ENS^2 and state-of-the-art models. $\Delta F/F_0$ calcium inputs and noise-matching inputs are shown with the ground truth (GT) on top. The predicted (PD) spike rates and corresponding spike events by various methods are shown below. Metrics on the right measure the performance on the corresponding neurons. Orange shaded areas represent regions of interest where discrepancies in predictions are significant among different methods.

(E) Comparison of neural networks adopted in ENS^2 and CASCADE.

(F) Comparison of the run time for training neural network models in ENS^2 and CASCADE.

(G) Comparison of run time spent for spike-rate inference among different algorithms. The run time is measured for each neuron and normalized to 1 min.

(H) Same as (G), but for spike-event inference. Colored circles present each neuron from all 26 datasets.

Error bars represent medians with 95% confidence intervals. Text denotes mean values of all neurons. Asterisks indicate significant difference using Friedman's test with Dunn's multiple comparisons (G) or two-sided Wilcoxon signed-rank test (F) and (H) between the indicated systems (****$p < 0.0001$). The run time was measured on a PC with an Intel Xeon E5 1630 v4 CPU and Nvidia GTX 1080 GPU.

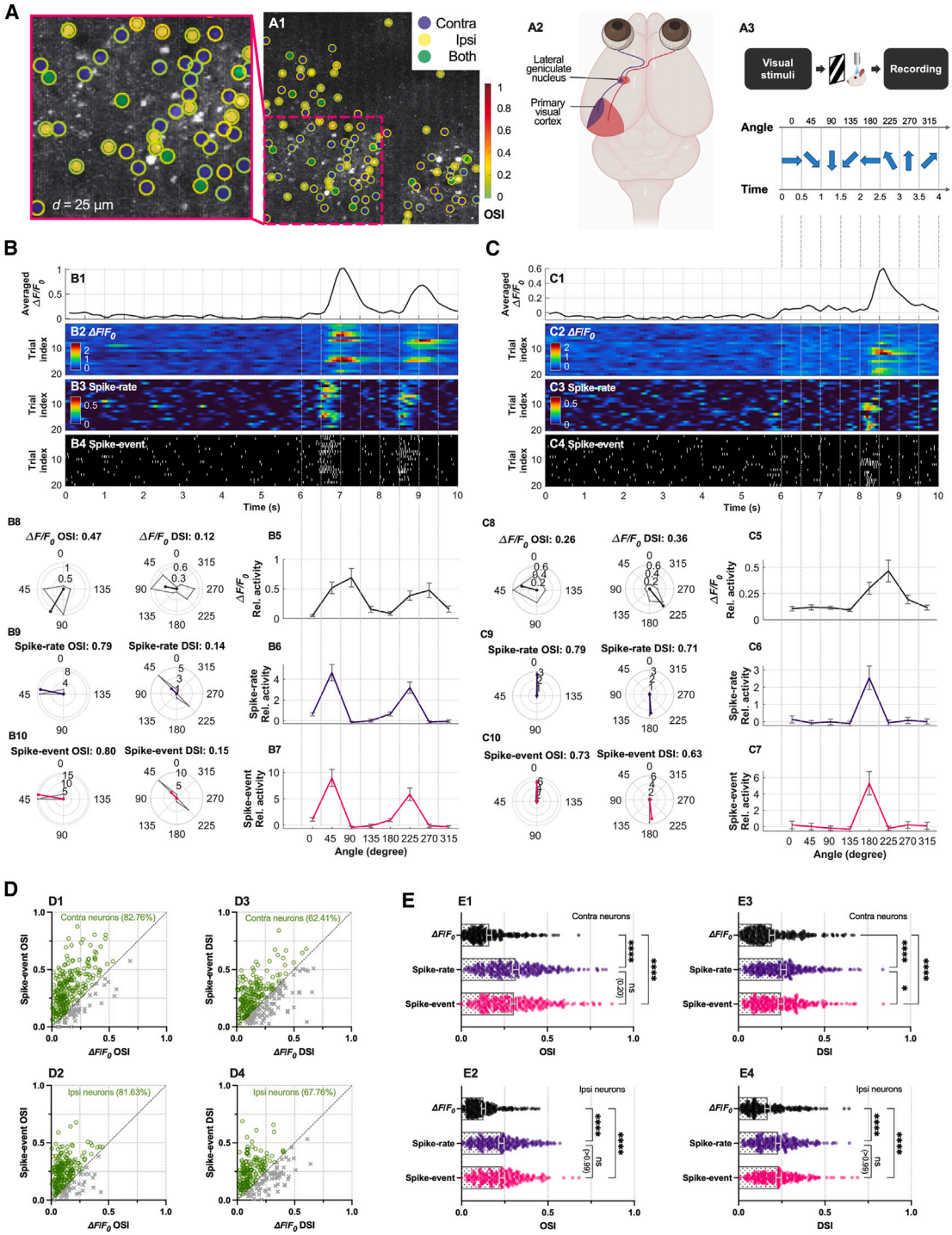


Figure 5. Spike inference with ENS^2 for calcium imaging data collected in primary visual cortex (V1) in a visual-stimulating experiment

(A) An example of image (A1) recorded from the binocular zone of the left V1 (A2) of mice subject to visual grating stimuli (A3). Visually responsive neurons are labeled in (A1) and are considered for further analyses (see STAR Methods). The inner color denotes the response type of the neurons. "Contra" and "ipsi" refers to the neurons that are responsive to contralateral (right) and ipsilateral (left) eye inputs, respectively. "Both" means the neurons are responsive to both sides of inputs. The outer color denotes the OSI computed from $\Delta F/F_0$ signal for that neuron.

(B and C) Examples of recorded calcium signals and the predicted spike rates and spike events by ENS^2 (B1–B4 and C1–C4). Tuning curves and selectivity indexes (B5–B10 and C5–C10) are computed based on three types of input (B2–B4 and C2–C4), respectively.

(legend continued on next page)

CASCADE system. However, both MLspike and OASIS performed poorly in this case, probably because of biased parameters during auto-calibrations. Figure 4C demonstrates an example with synthetic dyes, where the calcium dynamics are slow, and much lower frame rate at 7.8Hz, and hence the signal quality reduces. This slow dynamic caused shifted predictions in time on ENS^2, CASCADE, and MLspike. The low SNR caused MLspike to over-estimate, whereas OASIS failed to detect any spike in this segment. Nevertheless, the ENS^2 system still recovered the firing pattern the best among the four. In contrast, we have shown that these systems performed worse on inhibitory neurons (Figure S3), where the firing rate is generally much higher with bursting. Figure 4D shows the inference results on an inhibitory neuron. Considering the low frame rate and slow dynamic of calcium imaging, the ground-truth spiking activities would be extremely challenging to be recovered with such high firing rate. Nevertheless, predictions from ENS^2 and CASCADE resembled the temporal patterns of the ground truth much better than MLspike and OASIS, in that MLspike tended to under-estimate and OASIS tended to over-estimate. Together, we showed that our proposed ENS^2 systems could maintain robust inference capability under various conditions.

We would also like to point out that although our proposed ENS^2 is data driven, it is less computationally demanding than the previous method (e.g., CASCADE; Figure 4E). For a specific sampling rate (e.g., 60 Hz), series of noise matching models were trained in CASCADE to meet the need of different noise levels. Each of their noise-matching models consisted of five identical networks for ensemble learning to boost performance. In contrast, only a single network is required in our ENS^2 to predict data under each sampling rate for various conditions. This may benefit from our U-Net architecture design and model regularization. The U-Net architecture may provide higher capacity with a similar number of parameters. Proper regularization relieves over-fitting and reduces the need of ensemble from multiple models. As a result, the ENS^2 with U-Net requires 20,000 (around 18%) fewer trainable parameters, and only a maximum of 5.12 million data segments are fed for training, which is only 1/50 of that in CASCADE (Figure 4E). In fact, the ENS^2 system could perform at similar level with only 50,000 trainable parameters (around 30% of that in CASCADE) (Figure S1). In particular, training of ENS^2 on millions of samples during benchmarking was completed in around 2 min on average for each testing dataset, which was one order faster than CASCADE (Figure 4F). This will enable cost-effective re-training or fine-tuning when more paired datasets are available to improve our model further.

We also evaluated the efficiency of these systems during inference (Figures 4G and 4H). For spike-rate inference, our ENS^2 system took one order less time than CASCADE to complete for every 1 min of recording, and two orders less time than MLspike on average. The OASIS system showed the fastest spike-rate inference, despite compromised performance shown in our comparisons (Figures 3A–3D, S2C, and S2D). For spike-event inference, our ENS^2 system with the greedy estimation algorithm (see above; Figure 1B) took two orders less time to complete than MLspike. The run time was not measured for CASCADE as its built-in estimation algorithm took days to complete the benchmark for all datasets. The run time was not presented for OASIS either because the spike-event predictions were obtained by hard thresholding. Overall, we show that our ENS^2 demonstrated better performance and computational efficiency for inferring spikes from calcium data than the state-of-art methods.

Application to information encoding in V1

In the above benchmark, we showed that ENS^2 accomplished relatively high performance and high efficiency for inference in un-seen recordings. Next, we ask whether our system would provide additional insights to physiological observations *in vivo* as previous models did.[5,7,8] Here, we trained our full ENS^2 system with all 20 excitatory datasets available and then deployed it to un-seen calcium imaging data recorded from the V1 (Figures 5 and S4). We collected *in vivo* calcium fluorescence images with GCaMP6s indicators from V1 of mice that were shown to drift grating stimuli of four unique orientations that move in two opposing directions (eight directions total) (Figure 5A; see STAR Methods). Responsive neurons were selected, and their fluorescence signals were processed into calcium traces ($\Delta F/F_0$) for further analyses (Figures 5B1, 5B2, 5C1, and 5C2; see STAR Methods). We then used our ENS^2 system to predict the spike rates (Figures 5B3 and 5C3) and spike events (Figures 5B4 and 5C4) accordingly. We compared our analyses with these different inputs and verified whether the spike inference has any positive impact in understanding the information encoding in V1 than the raw $\Delta F/F_0$ trace alone. Here, two presentative neurons are shown in Figures 5B and 5C.

We first constructed the tuning curves (see STAR Methods) for each neuron (Figures 5B5–7 and 5C5–7). It was observed that the resultant tuning curves were broader when computed using $\Delta F/F_0$ (Figures 5B5 and 5C5) than using spike rate or spike event (Figures 5B6, 5B7, 5C6, and 5C7). In particular, the spike-rate/spike-event tuning curves exhibited sharpened preferred orientation for these cells. The broader tuning curve by $\Delta F/F_0$ was mainly due to the long "tail" (decaying edges) in $\Delta F/F_0$ signal after each peak (Figures 5B1 and 5C1) resulting from the slow dynamics of calcium indicators. Consequently, the long "tail" of $\Delta F/F_0$ may lead to a shift in the preferred orientation and broader tuning curves. In contrast, by predicting the spike rate/spike event with our ENS^2 system, we successfully removed these long tails in the signal. We further quantified the preferred orientations by computing the orientation selectivity index (OSI) and direction selectivity index (DSI) from the tuning curves for each neuron (see STAR Methods). The neuron in Figure 5B is a sample cell that exhibited high OSI but low DSI, whereas the neuron in Figure 5C is exhibiting high OSI and high DSI[54] (see STAR Methods). Our results show that OSI computed from $\Delta F/F_0$

(D and E) Comparisons of OSI/DSI computed from $\Delta F/F_0$ signal and after spike inference with ENS^2 for all 290 contra and 245 ipsi neurons considered. Each circle/cross/dot represents one recorded neuron.

Error bars represent medians with 95% confidence intervals. Asterisks indicate significant difference using Friedman's test with Dunn's multiple comparisons between the indicated predictions (****$p < 0.0001$).

was lower for both neurons in Figures 5B and 5C but was higher when computed from spike rate/spike event (from 0.47 to 0.79/0.80 and from 0.26 to 0.79/0.73, respectively). Similarly, DSI computed from $\Delta F/F_0$ was lower for the neuron in Figure 5C but increased when computed from spike rate/spike event (from 0.36 to 0.71/0.63). Figures S4A and S4B showed two neurons that appeared to have a preferred orientation at 0° but with a strong decaying tail in their $\Delta F/F_0$. Due to such up-shifted amplitude in $\Delta F/F_0$, the computed OSI or DSI tends to be higher. In contrast, the predicted spikes removed the up-shifting and resulted in a more clear-cut readout of the OSI/DSI measurement (such as in Figure S4B).

This observation is consistent for the neuron population we have recorded (290 contra and 245 ipsi neurons; see STAR Methods). Figures 5D1 and 5D2 showed that OSI computed with spike event was higher than that computed with $\Delta F/F_0$ in >80% of the cells (82.76% for contra neurons and 81.63% for ipsi neurons). Also, Figures 5D3 and 5D4 show that DSI computed with spike event was higher than that computed with $\Delta F/F_0$ in >60% of the cells (62.41% for contra neurons and 67.76% for ipsi neurons). Overall, the mean OSI/DSI for both contra and ipsi neurons was higher with spike event/spike rate than $\Delta F/F_0$ ($p < 0.0001$ for all cases; Figure 5E). From the results of all these cases, we believe that the OSI and DSI computed from spike rate/spike event predicted from our ENS^2 system would discriminate the response pattern of these neurons better than those derived from the original $\Delta F/F_0$.

It is also worth noting how the spike inference would benefit analyses of neurons that show weak responses or where the SNR is low (Figures S4C and S4D). In Figure S4C, the $\Delta F/F_0$ of this neuron varied over a range of only 0.2 such that the SNR was very low. The resultant noisy tuning curves suggested imprecise orientation or direction selectivity for this neuron. Instead, the spike inference increased the response SNR such that the tuning curve was much sharpened, showing preferred orientation at 0° and 180° and hence an increased OSI. On the contrary, in Figure S4D, the tuning curve from $\Delta F/F_0$ resulted in a sizable OSI (0.43), which may (falsely) suggest that this neuron has an orientation preference at 0°. Nevertheless, after spike inference with our ENS^2, the small peaks in the original $\Delta F/F_0$ signal were filtered out, resulting in negligible OSI and DSI. This suggested that this neuron in fact has very weak selectivity and may not be considered a real responsive cell. In this sense, the spike inference by our ENS^2 increased the SNR of the neuronal response to not only improve the sensitivity in detecting the orientation selectivity of the neurons but also to screen out some marginally responsive neurons.

Factors affecting inference performance

To understand further what contributed to good spike inference performance for data-driven methods, we investigated how the data itself affected the performance of these models. These insights may further facilitate data collection and preparation for improving data-driven models (e.g., our ENS^2).

Figures 6A–6D show the performance with individual dataset achieved by different configurations of models (the configurations are numbered horizontally in the same order as in Figures 2A–2D). The results show that performance indeed depended strongly on the dataset. For instance, regardless of the networks used, datasets 15 and 17 achieved notably worse vRD than other datasets (Figure 6B), and some datasets (e.g., 21 and 22) showed considerably worse ER than the others (Figure 6C). These were also observed when comparing our ENS^2 with state-of-the-art methods (Figures 3E–3H). We extracted a number of parameters from the calcium recording for each neuron, such as noise level, peak firing rate, frame rate, and calcium transient amplitude (see STAR Methods), and examined how they may affect the inference performance (Figures 6E–6H and S2I). Among these, it is shown that the transient amplitude and the ratio of transient amplitude by noise level (as a general index of SNR) of the dataset were the key predictors for the inference performance (including Corr, vRD, ER, and VPD). We also noticed that the original transient amplitudes in $\Delta F/F_0$ were of critical importance for inference, which related the number of spike events for a certain calcium indicator. A previous model-based study also testified that an accurate estimate of transient amplitude improved performance.[55] The transient amplitude indeed strongly depends on the calcium indicators' sensitivity. The frame rate also correlated with the inference performance significantly, probably because higher frame rate can capture the transient amplitude better.

We next investigated the preferred ways of supplying the training data for our system. We first examined this by training the model using two different subsets of the training data for each of the testing data, according to the types of calcium indicators. Here, "*All*" used all the possible training sets in training (same as the default leave-one-dataset-out protocol), whereas "*Same*" used only those training datasets with the same calcium indicator as the testing data. Note that these simulations were performed separately for datasets with excitatory and inhibitory neurons, respectively (see STAR Methods). Also note that some of the indicators comprise only one dataset (e.g., GCaMP5k) and hence they were not tested in the "*Same*" protocol. Figures 6J–6M and S2J show that "*Same*" did not perform better than "*All*" significantly in general. Similar observation was found in Rupprecht et al.,[38] where they reported that clustering the same calcium indicators for training showed no advantage in CASCADE. It seems that a generalized inference model ("*All*") is sufficiently good for the inference task than using multiple indicator-specific models ("*Same*"). The advantage of indicator-specific training was most prominent in datasets with GCaMP6f indicators (Figures 6I–6M, S2J, and S2K). This was probably because GCaMP6f is the dominating indicator in the whole benchmark database. Nevertheless, the difference remained quite small. As such, we recommend training the neural network-based model (e.g., our ENS^2 system) with all available paired data to exploit their generalization capability.

Because our ENS^2 is a data-driven model, we also wondered how much training data are needed for achieving good inference performance. We randomly sampled different numbers of segments from the total of over 20 h of available paired data from all excitatory datasets. When supplying all available paired data to the model, the total duration was approximately 20 × 96 h because the paired data were segmented with a step of one data point (see STAR Methods). Not surprisingly, the performance of inference increases with the amount of training data,

CellPress
OPEN ACCESS

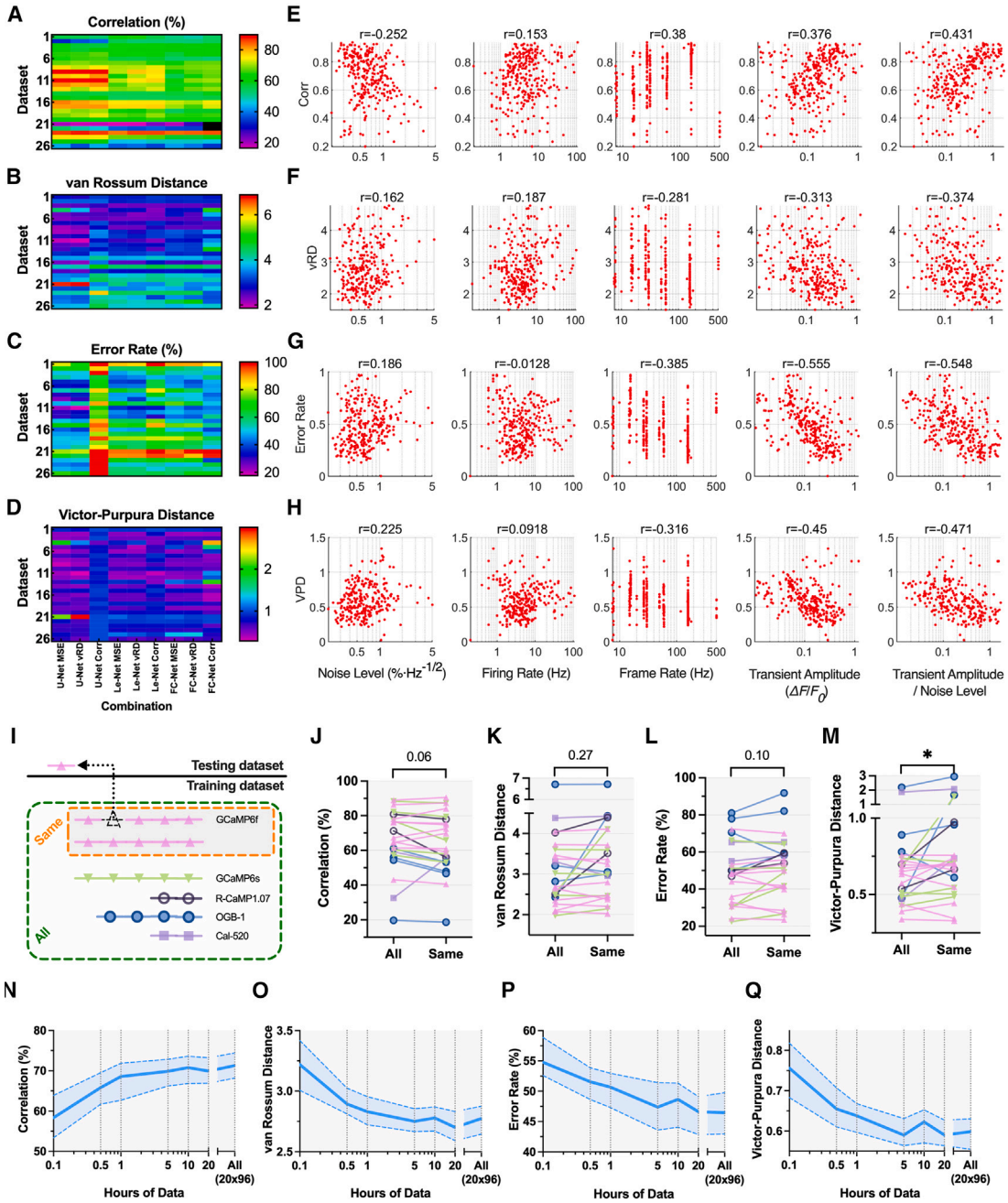


Figure 6. Effect of dataset properties, types of calcium indicator, and training data size on spike inference performance

(A–D) Performance with different configurations (see Figures 2A–2D) (x axis) on each dataset (y axis). Colormap shows the performance of the median neurons of each dataset.

(E–H) Spearman's rank correlation coefficients (r) between the five properties of each dataset and the corresponding spike inference performances. Red dots represent each neuron from all datasets.

(I) An example to illustrate how the division of datasets was made based on calcium indicator when a GCaMP6f dataset was regarded as testing dataset. When testing on an excitatory/inhibitory dataset, "All" refers to all the other 19/5 datasets, and "Same" refers to the datasets that also used GCaMP6f. Note that the simulations were performed separately for excitatory/inhibitory datasets.

(J–M) Performance of spike inference for different types of calcium indicator. Colored circles present the performance of median neurons of each testing dataset. Asterisks indicate significant difference using two-sided Wilcoxon signed-rank test between the indicated partitions (*$p < 0.05$).

(N–Q) Performance of spike inference with different length of training data. Shaded areas denote medians with 95% confidence intervals.

but it plateaued with roughly 5 h of paired data (Figures 6N–6Q, S2L, and S2M).

Several other factors may also have significant impact on the inference performance, such as sampling rate (resolution of prediction), size of smoothing window (for spike-rate prediction; see STAR Methods), and hyper-parameter of evaluation metric (e.g., ER window size; see STAR Methods). The comparisons are summarized in Figures 7A–7D. Figure 7A shows that the Corr increased consistently with larger smoothing windows. Similar observations can also be found in several recent studies.[35,36,56] This is because the GT spike rates convolved from the GT spike events with larger smoothing windows have smoother and broader patterns, which favored the measure of Corr. Figure 7F shows the GT spike rates obtained by convolving the GT spike events in Figure 7E with varying smoothing window sizes (25–200 ms) and their corresponding predictions. Apparently, the smoother and broader waveform of GT spike rate (with larger smoothing windows) simplified the prediction task, and it was easier to achieve a high Corr with such simpler and smoother PD spike-rate waveform. This was also true for the spike-rate evaluation with vRD and Error (Figures 7 and S2N). However, we argued that such resultant "better" performance (e.g., high Corr) would not always guarantee meaningful predictions as reflected in the PD spike events, because multiple GT spike events could be merged into a single peak of spike rate (Figures 7E and 7F). Instead, the temporal firing patterns could be better predicted with narrower smoothing windows. In contrast, spike-event predictions (VPD and ER; Figures 7C, 7D, and 7G) generally improved with higher sampling rates. This is quite reasonable because smaller bin sizes allow more precise estimation of spike events from spike-rate predictions. Moreover, when high sampling rates were used (e.g., 30 or 60 Hz), VPD and ER also reduced along with smoothing window sizes, indicating improved spike-event predictions. Here, the spike-event inference performance would possibly be restricted by the overly smoothed spike rates (e.g., Figure 7F). We also analyzed the effect of ER window sizes on ER evaluation (Figure 7G). As expected, smaller ER window sizes put higher demand on the evaluation of the algorithm and hence resulted in larger ER, which is similar to the findings reported in a previous study.[29]

Given these observations from our simulation results, we suggest that our ENS^2 system should be trained with inputs at sampling rate of 60 Hz (by re-sampling when necessary) with 25-ms smoothing windows for practical use (labeled with white dashed boxes in Figures 7A–7D, S2N, and S2O). Further increase in the sampling rate would cause computational overhead, whereas to reduce the smoothing window size further might be harmful to training neural networks with gradient descent. We also repeated our benchmark using Causal smoothing kernels as in CASCADE.[38] Nevertheless, we found that using Gaussian smoothing kernels generally outperforms Causal smoothing kernels ($p < 0.001$ for Corr/ER/Error; $p < 0.0001$ for vRD; $p = 0.42$ for VPD; data not shown). It is also worth noting that the CASCADE algorithm[38] was indeed benchmarked under 7.5 Hz with 200-ms smoothing windows. Figure S5 shows that our ENS^2 consistently outperformed the CASCADE algorithm at these settings for both spike-rate and spike-event predictions (Corr: $p = 0.0058$; vRD/error/bias: $p < 0.001$; ER/VPD: $p < 0.0001$) (also labeled in red dashed boxes in Figures 7A–7D, S2N, and S2O). These results support that our ENS^2 is a versatile and highly effective algorithm for spike inference from calcium signals.

DISCUSSION

In this work, we have developed a high-performance inference system (ENS^2) and showed its usefulness in inferring both spike rates and spike events. We found that networks with convolutional layers (e.g., U-Net and Le-Net) typically out-performed the other (e.g., FC-Net). This may be partly due to the regularization capability of the convolutional layers. In other words, it provides larger receptive fields (with context information such as calcium dynamics) with fewer trainable parameters and constrained kernel shapes. In contrast, it is quite intuitive for humans to examine the calcium segments fraction by fraction to identify spike events, just as sliding a kernel for convolution by the artificial neural networks. In fact, recent data-driven models (CASCADE,[38] S2S[39]) also used a network with convolutional layers. We speculate that the state-of-art performance of ENS^2 also benefits from the skip-connection structure and sequence-to-sequence prediction manner of our modified U-Net (Figure 1C). Recently, a 3D U-Net-based model has also been proposed to improve SNR in calcium images and facilitate calcium signal extraction.[57] In contrast, we revealed in our results that MSE loss could readily regulate the optimization of such models. Although Corr is indisputably a major evaluation metric for spike inference, we suggest that using Corr as the sole loss function for deep learning models (e.g., in S2S[39]) might not be ideal in real-world tasks. For example, the inferred spike rates may have the correct temporal pattern but are incorrect in the absolute amplitudes, hence the spike events cannot be recovered faithfully.

Although inferring spikes from calcium signals is a typical sequence-to-sequence translation task, recurrent neural networks (RNNs) (e.g., LSTM-based [long short-term memory] models[36]) may also be applied here. Nevertheless, RNN has to iteratively compute on each data point along time. Thus, the computational efficiency would be quite low compared with CNN-based models. A similar issue is also faced by the powerful Transformer-based models[58] designed for sequence-to-sequence translation, which have overly large numbers of parameters for online application of spike inference. On the contrary, in this work, we used deep convolutional networks with sequence-to-sequence translation ability (e.g., 1D U-Net) to retrieve local information in data segments for inference, and it showed state-of-art performance with desirable efficiency. To further improve the performance potentially, one may provide global information (e.g., statistical indicators such as global noise level) explicitly to the models. However, this requires sufficiently long recording for obtaining reliable statistics and would compromise the system efficiency when online application is needed.

Inhibitory neurons typically have higher instantaneous firing rates (Table S1), which complicated the spike inference task. Furthermore, the low frame rate and slow dynamics of calcium imaging may cause losses of even more details of the high-frequency spiking activities, leading to lower SNR in these neurons (e.g., Figure 4D). As a result, it should be expected that the

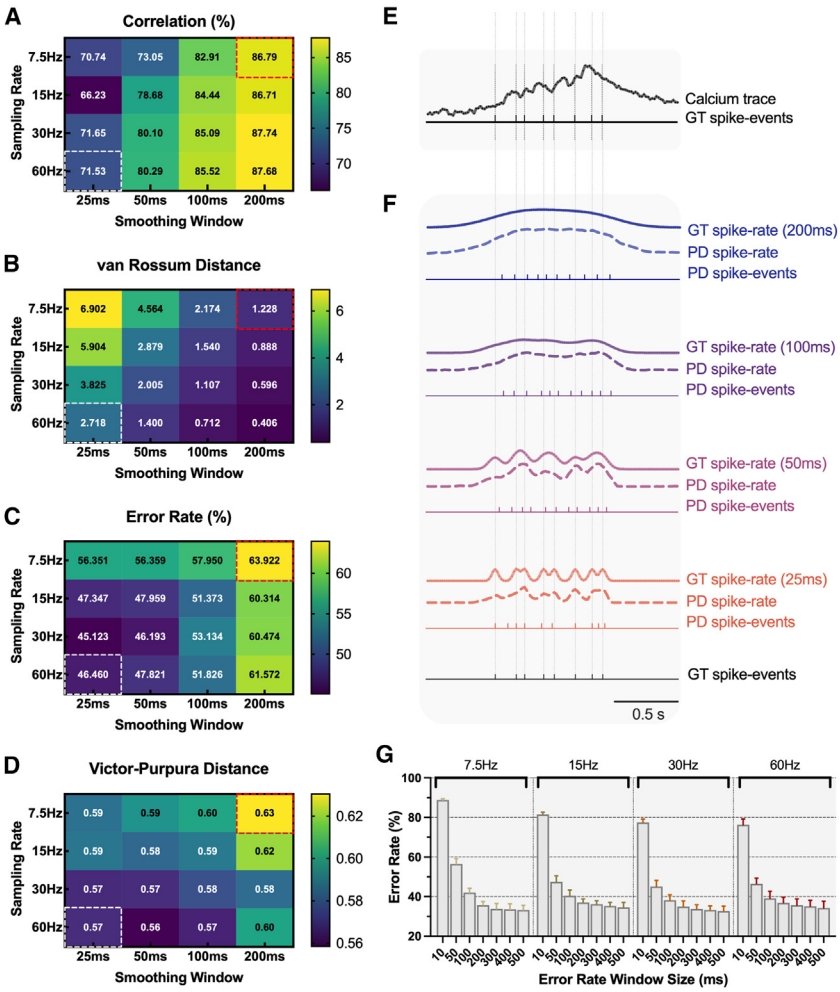


Figure 7. Effect of sampling rate, smoothing window size, and error rate window size on spike inference performance

(A–D) Performance in terms of correlation, van Rossum distance, error rate, and Victor-Purpura distance, respectively, when measured under different sampling rates and smoothing window sizes. Shown values represent median performance among all neurons. White and red squares denote the choice of sampling rate and smoothing window size adopted by ENS^2 and CASCADE, respectively.

(E) Example of calcium signals under 60 Hz with paired spike events.

(F) Examples of ground-truth (GT) spike rates convolved with different smoothing window sizes (from 200 to 25 ms). The resultant spike-rate and spike-event predictions (PD) are also shown.

(G) Performance of error rate when measured under different sampling rates and error rate window sizes. Error bars represent medians with 95% confidence intervals.

inference performance will be lower in these inhibitory neurons than the excitatory neurons. In contrast, the high firing rates also raise challenges to algorithms when discriminating the actual units of spikes in a single time step (time bin), because the transient amplitudes may change non-linearly as the number of spikes increases. It was shown that recovering every single spike from the calcium trace would be extremely difficult (Figure 4D). However, systems (e.g., ENS^2) that could properly predict the temporal firing patterns at a longer timescale for inhibitory neurons are still valuable tools for neuroscience research. Future work could explore how to incorporate the biophysical properties of these bursting patterns into the neural network, which will potentially improve the inference performance in the inhibitory neurons.[59]

For actual applications, data-driven methods (e.g., ENS^2 and CASCADE) are off the shelf for use without the need of further calibration, thanks to the generalization ability of neural networks, although CASCADE would need pre-training multiple versions to cater different noise levels. In contrast, for model-based methods (e.g., MLspike and OASIS), it is often necessary to define the parameters empirically or calibrate using some predefined algorithm for fresh recording, where ground-truth paired data are normally unavailable. The similar scenario was encountered in our leave-one-dataset-out benchmarking where the testing dataset was separated from the training set. Interestingly, we found that both tested model-based systems tend to under-estimate the spike rates and/or spike events (Figures S2D and S3F). The two data-driven systems appear to perform better on these un-seen data, which partly demonstrate their better generalization capability in this application.

Importantly, we have demonstrated that our spike inference algorithm could improve the analyses of real-world calcium data such as in the study of neuronal orientation preference in V1 (Figures 5 and S4). Our results demonstrate that our algorithm can help perform analyses with both high throughput (from calcium imaging) and high precision (from spiking activities) in the study of our brain.

Given that calcium recordings would be extremely noisy in certain experimental scenarios, we examined how the ENS^2 would perform in these conditions. We retrieved 50 artificial calcium recordings synthesized by the NAOMi generator[60] available from Rupprecht et al.[38] We then iteratively added white noise, with noise levels ranging from 1 to 15, onto them. We tested the existing ENS^2 system on these noisy data. The results are shown in Figure S6, where "NA" represents the noise-free synthetic data. The performance of ENS^2 indeed degraded as the data were becoming noisier. We showed some examples of these noisy calcium signals in Figure S6G, where the transient amplitudes were almost visually undistinguishable from the noise (e.g., noise level 15). We then retrained the ENS^2 model using the 20 benchmark datasets (as described in STAR Methods), but now white noise (up to noise level 15) was randomly added to the training calcium signal segments. We denote this new model trained with much noisier data as "Noise-Augmented ENS^2." We found that the Noise-Augmented ENS^2 generally performed

better than the original ENS^2 (Figure S6). This suggests that the designed U-Net architecture and model regularization help in generalizing to various noise levels in a single model. On one hand, this simplifies the inference pipeline, because a single model is sufficient to handle multiple noise levels, in contrary to some previous data-driven systems where multiple models were needed for various corresponding noise levels (e.g., CASCADE). On the other hand, this calibration-free and generalization capability of our data-driven system could also make it more convenient to use than the model-based systems.

We also tested the performance of the ENS^2 system in the presence of "pink noise" and "red noise," which simulate low-frequency biases or baseline drifting (Figure S7). Figure S7G shows some examples of synthetic calcium trace with added noise (white, pink, or red, with noise level of around 4). The performance of both original and Noise-Augmented ENS^2 significantly degraded in the pink or red noise (Figure S7). In fact, the Noise-Augmented model performed even worse than the original one when measured with vRD and VPD. Apparently, these low-frequency noises are more challenging to handle. It seems that they were less effectively captured in the U-Net architecture. Our results suggest that the low-frequency noise, such as irregular baseline drifts, may be better taken care of at the pre-processing stage.

Limitations of study

Several potential improvements for ENS^2 are outlined as follows. Because ENS^2 is a data-driven model, its performance would fluctuate upon different training data with various quality, diversity, and pre-processing procedures, etc. The benchmark database (Table S1) used in this work is essential to produce promising inference performance with the ENS^2 system, yet further improvement could also be made. We have re-sampled both the training data and testing data to 60 Hz for good-quality inference (Figure 7). The re-sampling was performed with a Fourier-based method (see STAR Methods), which may not provide notable additional information to our models. We suggest that adopting diffusion-based probabilistic models designed for biomedical time-series signal forecasting and imputation (e.g., Ho et al.,[61] Tashiro et al.,[62] and Alcaraz and Strodthoff[63]) may strengthen such re-sampling processes with temporal dependency. In contrast, it is possible that including more synthetic paired data (e.g., NAOMi[60]) and/or generative models (e.g., generative adversarial networks[64]) for data augmentation may further improve the generalization capability of our inference systems. As a preliminary study, we have shown above that introducing additive random noise to the training data artificially could improve the noise tolerance of our system. Lastly, our inference system is designed to be used in an end-to-end translation manner, where the inputs are $\Delta F/F_0$ calcium traces (see STAR Methods). In fact, extracting $\Delta F/F_0$ signals from calcium imaging is another non-trivial process, in addition to the spike inference task here. During this extraction process, the quality of the resultant $\Delta F/F_0$ calcium trace may limit and/or bias the inference of our system. We therefore envision that modifying our ENS^2 to directly take inputs from the source image-end (e.g., Pnevmatikakis[65]) and output to the spike-end may allow further exploitation and utilization of extra and unbiased information for spike inference.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Benchmark database
  - Data preparation: Re-sampling data
  - Data preparation: Pre-processing of inputs
  - Data preparation: Data segmentation
  - Evaluation of spike inference algorithms
  - Comparison to other state-of-arts models
  - *In vivo* experiments: Animal preparation and surgery
  - *In vivo* experiments: Two-photon calcium imaging
  - *In vivo* experiments: Visual stimulation protocols
  - *In vivo* experiments: Data processing and analysis
  - *In vivo* experiments: Calculation of tuning curves and selectivity indexes
  - Training and validation of Noise-Augmented ENS^2
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2023.100462.

ACKNOWLEDGMENTS

This work was supported by Research Grants Council of Hong Kong SAR (GRF 11104220 to C.T.; ECS 24117220 to J.P.K.I.), City University of Hong Kong, Hong Kong (Projects 7005645, 7005948, and 7020051 to C.T.), Lo Kwee-Seong Biomedical Research Fund (J.P.K.I.), and Faculty Innovation Awards (FIA2020/A/04) from the Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong (to J.P.K.I.).

AUTHOR CONTRIBUTIONS

C.T. conceived the study. Z.Z. designed the algorithms and performed analyses. J.P.K.I. and K.T. collected the *in vivo* calcium imaging data. J.P.K.I., K.T., and H.M.Y. pre-processed the data for spike inference and further analysis. M.S. provided expertise and inputs on dissecting V1 physiology and visual stimulation design. All of the authors contributed to interpreting the results and writing the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 31, 2022
Revised: February 21, 2023
Accepted: March 31, 2023
Published: April 24, 2023

REFERENCES

1. Neher, E., and Sakmann, B. (1976). Single-channel currents recorded from membrane of denervated frog muscle fibres. Nature *260*, 799–802. https://doi.org/10.1038/260799a0.
2. Hamill, O.P., Marty, A., Neher, E., Sakmann, B., and Sigworth, F.J. (1981). Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. Pflügers Archiv *391*, 85–100. https://doi.org/10.1007/BF00656997.
3. Spira, M.E., and Hai, A. (2013). Multi-electrode array technologies for neuroscience and cardiology. Nat. Nanotechnol. *8*, 83–94. https://doi.org/10.1038/nnano.2012.265.
4. Buzsáki, G. (2004). Large-scale recording of neuronal ensembles. Nat. Neurosci. *7*, 446–451. https://doi.org/10.1038/nn1233.
5. de Vries, S.E.J., Lecoq, J.A., Buice, M.A., Groblewski, P.A., Ocker, G.K., Oliver, M., Feng, D., Cain, N., Ledochowitsch, P., Millman, D., et al. (2020). A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. Nat. Neurosci. *23*, 138–151. https://doi.org/10.1038/s41593-019-0550-9.
6. Kerr, J.N.D., and Denk, W. (2008). Imaging in vivo: watching the brain in action. Nat. Rev. Neurosci. *9*, 195–205. https://doi.org/10.1038/nrn2338.
7. Wilson, N.R., Runyan, C.A., Wang, F.L., and Sur, M. (2012). Division and subtraction by distinct cortical inhibitory networks in vivo. Nature *488*, 343–348. https://doi.org/10.1038/nature11347.
8. Rikhye, R.V., and Sur, M. (2015). Spatial correlations in natural scenes modulate response reliability in mouse visual cortex. J. Neurosci. *35*, 14661–14680. https://doi.org/10.1523/JNEUROSCI.1660-15.2015.
9. Giovannucci, A., Badura, A., Deverett, B., Najafi, F., Pereira, T.D., Gao, Z., Ozden, I., Kloth, A.D., Pnevmatikakis, E., Paninski, L., et al. (2017). Cerebellar granule cells acquire a widespread predictive feedback signal during motor learning. Nat. Neurosci. *20*, 727–734. https://doi.org/10.1038/nn.4531.
10. Knogler, L.D., Markov, D.A., Dragomir, E.I., Štih, V., and Portugues, R. (2017). Sensorimotor representations in cerebellar granule cells in larval zebrafish are dense, spatially organized, and non-temporally patterned. Curr. Biol. *27*, 1288–1302. https://doi.org/10.1016/j.cub.2017.03.029.
11. Wagner, M.J., Kim, T.H., Savall, J., Schnitzer, M.J., and Luo, L. (2017). Cerebellar granule cells encode the expectation of reward. Nature *544*, 96–100. https://doi.org/10.1038/nature21726.
12. El-Boustani, S., Ip, J.P.K., Breton-Provencher, V., Knott, G.W., Okuno, H., Bito, H., and Sur, M. (2018). Locally coordinated synaptic plasticity of visual cortex neurons in vivo. Science *360*, 1349–1354. https://doi.org/10.1126/science.aao0862.
13. Akerboom, J., Chen, T.-W., Wardill, T.J., Tian, L., Marvin, J.S., Mutlu, S., Calderón, N.C., Esposti, F., Borghuis, B.G., Sun, X.R., et al. (2012). Optimization of a GCaMP calcium indicator for neural activity imaging. J. Neurosci. *32*, 13819–13840. https://doi.org/10.1523/JNEUROSCI.2601-12.2012.
14. Chen, T.-W., Wardill, T.J., Sun, Y., Pulver, S.R., Renninger, S.L., Baohan, A., Schreiter, E.R., Kerr, R.A., Orger, M.B., Jayaraman, V., et al. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature *499*, 295–300. https://doi.org/10.1038/nature12354.
15. Tada, M., Takeuchi, A., Hashizume, M., Kitamura, K., and Kano, M. (2014). A highly sensitive fluorescent indicator dye for calcium imaging of neural activity in vitro and in vivo. Eur. J. Neurosci. *39*, 1720–1728. https://doi.org/10.1111/ejn.12476.
16. Dana, H., Mohar, B., Sun, Y., Narayan, S., Gordus, A., Hasseman, J.P., Tsegaye, G., Holt, G.T., Hu, A., Walpita, D., et al. (2016). Sensitive red protein calcium indicators for imaging neural activity. Elife *5*, e12727. https://doi.org/10.7554/eLife.12727.
17. Bethge, P., Carta, S., Lorenzo, D.A., Egolf, L., Goniotaki, D., Madisen, L., Voigt, F.F., Chen, J.L., Schneider, B., Ohkura, M., et al. (2017). An R-CaMP1.07 reporter mouse for cell-type-specific expression of a sensitive red fluorescent calcium indicator. PLoS One *12*, e0179460. https://doi.org/10.1371/journal.pone.0179460.
18. Denk, W., Strickler, J.H., and Webb, W.W. (1990). Two-photon laser scanning fluorescence microscopy. Science *248*, 73–76. https://doi.org/10.1126/science.2321027.
19. Stosiek, C., Garaschuk, O., Holthoff, K., and Konnerth, A. (2003). In vivo two-photon calcium imaging of neuronal networks. Proc. Natl. Acad. Sci. USA *100*, 7319–7324. https://doi.org/10.1073/pnas.1232232100.
20. Sofroniew, N.J., Flickinger, D., King, J., and Svoboda, K. (2016). A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. Elife *5*, e14472. https://doi.org/10.7554/eLife.14472.
21. Grienberger, C., and Konnerth, A. (2012). Imaging calcium in neurons. Neuron *73*, 862–885. https://doi.org/10.1016/j.neuron.2012.02.011.
22. Kerr, J.N.D., Greenberg, D., and Helmchen, F. (2005). Imaging input and output of neocortical networks in vivo. Proc. Natl. Acad. Sci. USA *102*, 14063–14068. https://doi.org/10.1073/pnas.0506029102.
23. Yaksi, E., and Friedrich, R.W. (2006). Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca2+ imaging. Nat. Methods *3*, 377–383. https://doi.org/10.1038/nmeth874.
24. Greenberg, D.S., Houweling, A.R., and Kerr, J.N.D. (2008). Population imaging of ongoing neuronal activity in the visual cortex of awake rats. Nat. Neurosci. *11*, 749–751. https://doi.org/10.1038/nn.2140.
25. Grewe, B.F., Langer, D., Kasper, H., Kampa, B.M., and Helmchen, F. (2010). High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. Nat. Methods *7*, 399–405. https://doi.org/10.1038/nmeth.1453.
26. Vogelstein, J.T., Packer, A.M., Machado, T.A., Sippy, T., Babadi, B., Yuste, R., and Paninski, L. (2010). Fast nonnegative deconvolution for spike train inference from population calcium imaging. J. Neurophysiol. *104*, 3691–3704. https://doi.org/10.1152/jn.01073.2009.
27. Oñativia, J., Schultz, S.R., and Dragotti, P.L. (2013). A finite rate of innovation algorithm for fast and accurate spike detection from two-photon calcium imaging. J. Neural. Eng. *10*, 046017.
28. Pnevmatikakis, E.A., Merel, J., Pakman, A., and Paninski, L. (2013). Bayesian spike inference from calcium imaging data *2013*, 349–353.
29. Deneux, T., Kaszas, A., Szalay, G., Katona, G., Lakner, T., Grinvald, A., Rózsa, B., and Vanzetta, I. (2016). Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo. Nat. Commun. *7*, 12190. https://doi.org/10.1038/ncomms12190.
30. Friedrich, J., and Paninski, L. (2016). Fast active set methods for online spike inference from calcium imaging. Adv. Neural Inf. Process. Syst. *29*.
31. Pnevmatikakis, E.A., Soudry, D., Gao, Y., Machado, T.A., Merel, J., Pfau, D., Reardon, T., Mu, Y., Lacefield, C., Yang, W., et al. (2016). Simultaneous denoising, deconvolution, and demixing of calcium imaging data. Neuron *89*, 285–299. https://doi.org/10.1016/j.neuron.2015.11.037.
32. Friedrich, J., Zhou, P., and Paninski, L. (2017). Fast online deconvolution of calcium imaging data. PLoS Comput. Biol. *13*, e1005423. https://doi.org/10.1371/journal.pcbi.1005423.
33. Pachitariu, M., Stringer, C., and Harris, K.D. (2018). Robustness of spike deconvolution for neuronal calcium imaging. J. Neurosci. *38*, 7976–7985. https://doi.org/10.1523/JNEUROSCI.3339-17.2018.
34. Jewell, S.W., Hocking, T.D., Fearnhead, P., and Witten, D.M. (2020). Fast nonconvex deconvolution of calcium imaging data. Biostatistics *21*, 709–726. https://doi.org/10.1093/biostatistics/kxy083.
35. Theis, L., Berens, P., Froudarakis, E., Reimer, J., Román Rosón, M., Baden, T., Euler, T., Tolias, A.S., and Bethge, M. (2016). Benchmarking spike rate inference in population calcium imaging. Neuron *90*, 471–482. https://doi.org/10.1016/j.neuron.2016.04.014.
36. Berens, P., Freeman, J., Deneux, T., Chenkov, N., McColgan, T., Speiser, A., Macke, J.H., Turaga, S.C., Mineault, P., Rupprecht, P., et al. (2018). Community-based benchmarking improves spike rate inference from

two-photon calcium imaging data. PLoS Comput. Biol. *14*, e1006157. https://doi.org/10.1371/journal.pcbi.1006157.

37. Hoang, H., Sato, M.-a., Shinomoto, S., Tsutsumi, S., Hashizume, M., Ishikawa, T., Kano, M., Ikegaya, Y., Kitamura, K., Kawato, M., and Toyama, K. (2020). Improved hyperacuity estimation of spike timing from calcium imaging. Sci. Rep. *10*, 17844. https://doi.org/10.1038/s41598-020-74672-y.
38. Rupprecht, P., Carta, S., Hoffmann, A., Echizen, M., Blot, A., Kwan, A.C., Dan, Y., Hofer, S.B., Kitamura, K., Helmchen, F., and Friedrich, R.W. (2021). A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging. Nat. Neurosci. *24*, 1324–1337. https://doi.org/10.1038/s41593-021-00895-5.
39. Sebastian, J., Sur, M., Murthy, H.A., and Magimai-Doss, M. (2021). Signal-to-signal neural networks for improved spike estimation from calcium imaging data. PLoS Comput. Biol. *17*, e1007921. https://doi.org/10.1371/journal.pcbi.1007921.
40. Tsutsumi, S., Yamazaki, M., Miyazaki, T., Watanabe, M., Sakimura, K., Kano, M., and Kitamura, K. (2015). Structure-function relationships between aldolase C/zebrin II expression and complex spike synchrony in the cerebellum. J. Neurosci. *35*, 843–852. https://doi.org/10.1523/JNEUROSCI.2170-14.2015.
41. Sebastian, J., Kumar, M.G., Viraraghavan, V.S., Sur, M., and Murthy, H.A. (2019). Spike estimation from fluorescence signals using high-resolution property of group delay. IEEE Trans. Signal Process. *67*, 2923–2936. https://doi.org/10.1109/TSP.2019.2908913.
42. Zhai, X., and Tin, C. (2018). Automated ECG classification using dual heartbeat coupling based on convolutional neural network. IEEE Access *6*, 27465–27472. https://doi.org/10.1109/access.2018.2833841.
43. Zhai, X., Zhou, Z., and Tin, C. (2020). Semi-supervised learning for ECG classification without patient-specific labeled data. Expert Syst. Appl. *158*, 113411. https://doi.org/10.1016/j.eswa.2020.113411.
44. Zhou, Z., Zhai, X., and Tin, C. (2021). Fully automatic electrocardiogram classification system based on generative adversarial network with auxiliary classifier. Expert Syst. Appl. *174*, 114809. https://doi.org/10.1016/j.eswa.2021.114809.
45. Zhai, X., Jelfs, B., Chan, R.H.M., and Tin, C. (2017). Self-recalibrating surface EMG pattern recognition for neuroprosthesis control based on convolutional neural network. Front. Neurosci. *11*, 379. https://doi.org/10.3389/fnins.2017.00379.
46. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation., N. Navab, J. Hornegger, W.M. Wells, and A.F. Frangi, eds. (Springer International Publishing), pp. 234–241.
47. Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proc. IEEE *86*, 2278–2324. https://doi.org/10.1109/5.726791.
48. Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: the missing ingredient for fast stylization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1607.08022.
49. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. *15*, 1929–1958.
50. van Rossum, M.C. (2001). A novel spike distance. Neural Comput. *13*, 751–763. https://doi.org/10.1162/089976601300014321.
51. Kingma, D.P., and Ba, J. (2015). Adam: a method for stochastic optimization. 3rd International Conference on Learning Representations (ICLR).
52. Pachitariu, M., Stringer, C., Dipoppa, M., Schröder, S., Rossi, L.F., Dalgleish, H., Carandini, M., and Harris, K.D. (2017). Suite2p: beyond 10,000 neurons with standard two-photon microscopy. Preprint at bioRxiv. https://doi.org/10.1101/061507.
53. Giovannucci, A., Friedrich, J., Gunn, P., Kalfon, J., Brown, B.L., Koay, S.A., Taxidis, J., Najafi, F., Gauthier, J.L., Zhou, P., et al. (2019). CaImAn an open source tool for scalable calcium imaging data analysis. Elife *8*, e38173. https://doi.org/10.7554/eLife.38173.
54. Mazurek, M., Kager, M., and Van Hooser, S.D. (2014). Robust quantification of orientation selectivity and direction selectivity. Front. Neural Circ. *8*, 92. https://doi.org/10.3389/fncir.2014.00092.
55. Éltes, T., Szoboszlay, M., Kerti-Szigeti, K., and Nusser, Z. (2019). Improved spike inference accuracy by estimating the peak amplitude of unitary [Ca2+] transients in weakly GCaMP6f-expressing hippocampal pyramidal cells. J. Physiol. *597*, 2925–2947. https://doi.org/10.1113/jp277681.
56. Stringer, C., and Pachitariu, M. (2019). Computational processing of neural recordings from calcium imaging data. Curr. Opin. Neurobiol. *55*, 22–31. https://doi.org/10.1016/j.conb.2018.11.005.
57. Li, X., Zhang, G., Wu, J., Zhang, Y., Zhao, Z., Lin, X., Qiao, H., Xie, H., Wang, H., Fang, L., and Dai, Q. (2020). Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised learning. Preprint at bioRxiv. https://doi.org/10.1101/2020.11.16.383984.
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. Preprint at arXiv. https://doi.org/10.48550/arXiv.1706.03762.
59. Rahmati, V., Kirmse, K., Marković, D., Holthoff, K., and Kiebel, S.J. (2016). Inferring neuronal dynamics from calcium imaging data using biophysical models and bayesian inference. PLoS Comput. Biol. *12*, e1004736. https://doi.org/10.1371/journal.pcbi.1004736.
60. Charles, A.S., Song, A., Gauthier, J.L., Pillow, J.W., and Tank, D.W. (2019). Neural anatomy and optical microscopy (NAOMi) simulation for evaluating calcium imaging methods. Preprint at bioRxiv. https://doi.org/10.1101/726174.
61. Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. *33*, 6840–6851.
62. Tashiro, Y., Song, J., Song, Y., and Ermon, S. (2021). CSDI: conditional score-based diffusion models for probabilistic time series imputation. Adv. Neural Inf. Process. Syst. *34*, 24804–24816.
63. Alcaraz, J.M.L., and Strodthoff, N. (2022). Diffusion-based time series imputation and forecasting with structured state space models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2208.09399.
64. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems, *Volume 2* (MIT Press).
65. Pnevmatikakis, E.A. (2019). Analysis pipelines for calcium imaging data. Curr. Opin. Neurobiol. *55*, 15–21.
66. Huang, L., Knoblich, U., Ledochowitsch, P., Lecoq, J., Reid, R.C., de Vries, S.E.J., Buice, M.A., Murphy, G.J., Waters, J., Koch, C., et al. (2020). Relationship between simultaneously recorded spiking activity and fluorescence signal in GCaMP6 transgenic mice. Preprint at bioRxiv. https://doi.org/10.1101/788802.
67. Khan, A.G., Poort, J., Chadwick, A., Blot, A., Sahani, M., Mrsic-Flogel, T.D., and Hofer, S.B. (2018). Distinct learning-induced changes in stimulus selectivity and interactions of GABAergic interneuron classes in visual cortex. Nat. Neurosci. *21*, 851–859. https://doi.org/10.1038/s41593-018-0143-z.
68. Schoenfeld, G., Carta, S., Rupprecht, P., Ayaz, A., and Helmchen, F. (2021). In vivo calcium imaging of CA3 pyramidal neuron populations in adult mouse hippocampus. Preprint at bioRxiv. https://doi.org/10.1101/2021.01.21.427642.
69. Kwan, A.C., and Dan, Y. (2012). Dissection of cortical microcircuits by single-neuron stimulation in vivo. Curr. Biol. *22*, 1459–1467.
70. Victor, J.D., and Purpura, K.P. (1996). Nature and precision of temporal coding in visual cortex: a metric-space analysis. J. Neurophysiol. *76*, 1310–1326. https://doi.org/10.1152/jn.1996.76.2.1310.
71. Stoyanov, M., Gunzburger, M., and Burkardt, J. (2011). Pink noise, 1/f α noise, and their effect on solutions of differential equations. Int. J. Uncertain. Quantification *1*, 257–278.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Benchmark database	Rupprecht et al.[38]	https://doi.org/10.1038/s41593-021-00895-5
Experimental models: Organisms/strains		
Ai148 mice	The Jackson Laboratory	RRID:IMSR_JAX:030328
Software and algorithms		
ENS^2 (this work)	https://github.com/TinLab/ENS2	https://doi.org/10.5281/zenodo.7787553
CASCADE[38]	https://github.com/HelmchenLabSoftware/Cascade	https://doi.org/10.5281/zenodo.5477429
MLspike[29]	https://github.com/MLspike	N/A
OASIS[32]	https://github.com/j-friedrich/OASIS	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Chung Tin (chungtin@cityu.edu.hk).

Materials availability

This study did not generate new materials.

Data and code availability

- This paper analyzes existing, publicly available data as of the date of publication. DOI of the source of benchmark database is listed in the key resources table. *In vivo* calcium imaging data from mice V1 reported in this paper will be shared by the lead contact upon request.
- All original code has been deposited at https://github.com/tinlab/ens2 and the linked repositories therein, and is publicly available as of the date of publication. The DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ai148 mice (Jackson Lab strain #030328) were crossed with CamKII-cre mice to express GCaMP6s in excitatory neurons. Postnatal day (P) 25 mice (both male and female) were used in our experiments. All procedures were conducted under protocols approved by the Massachusetts Institute of Technology's Animal Care and Use Committee and conformed to US National Institutes of Health (NIH) guidelines.

METHOD DETAILS

Benchmark database

In this study, we used the publicly available datasets containing both calcium imaging signals and simultaneously recorded electrophysiological signals from excitatory neurons[13–17,35,38,66–68] and inhibitory neurons.[38,67,69] For benchmarking and algorithm development purposes, they were recently compiled by[38] into an extensive database with 27 datasets. Specifically, we adopted dataset #2 to #27 following[38] for a fair comparison, and they are labeled as dataset 1 to 26 in this study as shown in Table S1. Specifically, dataset 1 to 20 cover imaging of excitatory neurons from eight different kinds of calcium indicators, a wide range of frame rates (7.7Hz–500Hz), and various peak firing rates (1.18Hz–12.67Hz, averaged on each dataset). Over 20 h of paired ground truth data (calcium signals and spike-events) were recorded from a total of 229 neurons of either mouse or zebrafish brains. On the other hand, dataset 21 to 26 contain inhibitory neurons with much higher peak firing rates (up to 57.31Hz). There is a total of over 15 h of paired data from 16 *in vivo* and 41 *in vitro* inhibitory neurons.

In each dataset, calcium signals are provided as the percentage changes of fluorescence amplitude against baseline ($\Delta F/F_0$), while individual time stamps label spike-events. We also computed the noise-levels as defined in Rupprecht et al.[38] and listed them in Table S1. In brief, the noise-levels ν are computed by,

$$\nu = \frac{Median\{|DFF_{t+1} - DFF_t|\}}{\sqrt{fs}}$$

where DFF_t represents the fluorescence amplitude $\Delta F/F_0$ at any given time t; and fs is the sampling rate of the given calcium recording. This formula computes the median fluctuation value from a whole recording and scales it with $\sqrt{fs}$, so that it is quantitatively comparable across datasets.[38] Furthermore, we presented the increase in $\Delta F/F_0$ induced by one action potential (calcium transient amplitude) for each dataset. The transient amplitude is computed using the averaged calcium kernel, which was extracted from paired ground truth data using the deconvolution function with regularized filter ("deconvreg") in MATLAB. We computed the approximate instantaneous firing rates of a neuron using a 5 s sliding window with steps of 1/60 s (or 1 data point). The 95% quantile values of these computed firing rates were defined as the peak firing rate of this neuron. The values are shown as mean ±1 standard deviation in Table S1.

Data preparation: Re-sampling data

To develop and validate the spike inference algorithms, we first re-sampled the input data (both training set and testing set) of different frame rates to the same sampling rates. In this work, we referred to the original frequencies where calcium signals were captured as *frame rates*, and the re-sampled frequencies as *sampling rates*. Given that most of the datasets were captured with frame rates not higher than 60Hz (Table S1), we re-sampled all calcium signals to 60Hz. All the inference systems were then benchmarked under this same sampling rate. We also tested our system under 7.5Hz as suggested by CASCADE.[38] The re-sampling is performed with the "resample" function of SciPy. The impact of sampling rates on inference results is discussed in this work.

Data preparation: Pre-processing of inputs

We used the original $\Delta F/F_0$ calcium inputs for our system (where only re-sampling is performed). The training target (expected outputs from the systems) are prepared as below. For a pre-defined sampling rate (e.g. 60Hz), raw time stamps of ground truth spike (spike-events) are re-allocated into their corresponding time bins. We can then compute the sequence of spike counts by counting the total firing events in each time bin. The sequences are then smoothed with Gaussian filters to facilitate gradient descent. The smoothing window size τ for the Gaussian kernels was set to 25ms, which produces the optimal spike-event predictions with high temporal resolution in general. The selection of smoothing window size for deep learning based systems is also carefully studied. The convolved spike counts are denoted as "spike-rate" in this work.

Data preparation: Data segmentation

To train the neural networks properly, paired sequences of calcium signals and spike-rates were segmented with a moving step of 1 data point (Figure 1A). The length of each segment was set to 96 data points. In the case of sampling rate of 60Hz, a total of ~4 million segments of paired data were obtained for training.

Evaluation of spike inference algorithms

How to reliably assess the performance of the spike inference tasks remains an open topic, where a single evaluation metric could be biased in certain aspects.[35,36,38,39] In this regard, recent studies proposed to employ multiple metrics to supplement each other.[29,35–39] In this work, we used four metrics to examine spike-rates prediction and two others for spike-events prediction.

Firstly, Pearson correlation coefficient (Corr) is used as the primary metric for comparing similarities of spike rates as follow:

$$Corr_{GT,PD} = \frac{E[(GT - \mu_{GT})(PD - \mu_{PD})]}{\sigma_{GT}\sigma_{PD}} \quad \text{(Equation 4)}$$

where GT and PD stand for ground truth and prediction, respectively. Secondly, we use the van Rossum distance (vRD)[50] for the evaluation of spike rates prediction:

$$vRD_{GT,PD} = \sqrt{\frac{1}{\tau}\frac{\int [GT(t) - PD(t)]^2 dt}{\int [GT(t)]^2 dt}} \quad \text{(Equation 5)}$$

where the time constant τ is the normalizing factor (smoothing window size) for smoothing spike-events into spike-rates (e.g., τ = 0.025*s* for our proposed system). Moreover, Error and Bias proposed in[38] are also used to evaluate spike-rates:

$$Error = \frac{\int |PD - GT| dt}{\int GT dt} \quad \text{(Equation 6)}$$

$$Bias = \frac{\int (PD - GT)dt}{\int GTdt} \quad \text{(Equation 7)}$$

On the other hand, for measuring spike-event prediction, we adopt the Victor-Purpura distance (VPD).[70] It is defined as the minimal cost to transform the PD spike-events to the GT spike-events. The cost for either inserting or deleting a spike equals 1, while shifting a spike by Δt costs $q|\Delta t|$. We use the default value $q = 1$ in this work. To make comparison across different datasets, we present the VPD as the minimal total cost divided by the total number of GT spikes.

Lastly, we compute the error rate (ER) as below,[29,37,55] which measures the F_1 score of the predicted spike-events:

$$ER_{GT,PD} = 1 - F_1 = 1 - 2\frac{sensitivity \times precision}{sensitivity + precision} \quad \text{(Equation 8)}$$

$$sensitivity = \frac{true\ positive}{true\ positive + false\ negative} \quad \text{(Equation 9)}$$

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \quad \text{(Equation 10)}$$

The GT spike-events and PD spike-events are matched based on their VPD. Here, a spike is said to be correctly predicted if it co-exists with its real counterpart within a time window of 50ms (defined as the ER window size). This time window is one order smaller than that used in previous study,[29] suggesting a much more stringent assessment of model performance in this study. We also examined the effect of ER window sizes in this work.

Comparison to other state-of-arts models

Our ENS^2 system was compared against the state-of-the-art algorithms, including the data-driven method, CASCADE[38] and the model-based methods, MLspike[29] and OASIS.[32] All of these followed the leave-one-dataset-out protocol.

For the CASCADE algorithm, we followed the training protocol as described.[38] For each dataset, the "noise matching inputs" were obtained from CASCADE to reproduce results with the algorithm (e.g. artificial noise is added to the 19 training datasets to match the noise-level of the testing excitatory neurons, and vice versa for inhibitory neurons). Five identical models were trained separately for 10 epochs. The averaged outputs of these five models were regarded as the final spike-rate predictions. When testing under 60Hz sampling rate, the smoothing kernel size was reduced from 0.2s to 0.025s. Spike-event predictions are estimated using a Monte-Carlo importance sampling based algorithm in CASCADE.[38] All hyper-parameters are kept intact as in the CASCADE algorithms.

For the MLspike algorithm,[29] original $\Delta F/F_0$ calcium inputs are used. For datasets with OGB synthetic dyes, we set saturation $\gamma = 0.091$. For datasets with GECIs, we used the full physiological model version of MLspike with parameters modeling saturation, Hill exponent, c0, and rise time as described in Figure S6A of Deneux et al.[29] For the remaining datasets, we used the polynomial nonlinearity modeling in MLspike with coefficient $[p_2, p_3] = [1.0, 0.0]$. The values of the model parameters, A (transient amplitude), τ (calcium decay time constant), and σ (noise amplitude) were obtained using its built-in auto-calibration algorithm, since manual calibration on fresh recordings without ground truth is also challenging in actual application (see discussion). In case that the auto-calibration failed, we supplied the parameters of A and τ manually (following the methods shown in Figure S6A of Deneux et al.[29]). We have tested that MLspike performed worse when the two parameters were set manually. For a fair comparison, in addition to the spike-event outputs from MLspike, we also obtain its native *spikest_prob* output for evaluating its performance in spike-rate inference.

For the OASIS algorithm,[32] we used the L_1-regularized version by calling the *deconvolve* function. Original $\Delta F/F_0$ calcium traces were fed as inputs. We set *optimize_g* = 5 to auto-calibrate the parameter g. We also repeated the benchmark with L_0-regularization and/or by setting other g values, and the overall performance were similar. The hard thresholds for obtaining discrete spike-event outputs were computed as 55% (slightly over one-half as described in OASIS) of the real calcium transient amplitude of each neuron, following the suggestion in OASIS.[32] Note that without ground truth paired data, the hard thresholds will need to be estimated iteratively.

In vivo experiments: Animal preparation and surgery

Ai148 mice (Jackson Lab strain #030328) were crossed with CamKII-cre mice to express GCaMP6s in excitatory neurons. Animal surgery was described previously.[12] In brief, postnatal day (P) 25 mice were anesthetized with 3% isoflurane and confined by

stereotaxic frame. Scalp was sterilized and removed for the cranial window surgery. Skull above the left binocular visual cortex (Figure 5A3) was replaced by a 3mm/5mm stacked circular glass coverslip to ensure transparency for imaging. A tailor-made head-plate was fixed on the skull with Metabond adhesive cement.

In vivo experiments: Two-photon calcium imaging

The imaging process was performed by two-photon system with awake and head-fixed mice. The mice were allowed to recover for at least 3 days after the craniotomy, followed by a habituation on head-fixation. Prairie Ultima with a Spectra Physics Mai-Tai Deep See laser two-photon system (Prairie Technologies) was used for imaging. 20× Olympus objective lens was used for functional imaging. The calcium signals of neurons from layer $^{2}/_{3}$ were visualized at 920nm laser wavelength with acquisition frame rate around 7.6Hz (averaging from 4 consecutive frames around 30Hz).

In vivo experiments: Visual stimulation protocols

The visual stimulus was delivered from Psychtoolbox-3. A computer was connected to a 10-inch 1080p LCD monitor for display. Drifting gratings were used to stimulate the visual cortex (Figure 5A2-3). Each trial of display lasts for 10 s and repeated for 20 trials per stimulus, resulting in imaging session of 200 s. At the beginning of each trial, 6 s of gray screen is presented to the animals (defined as the *resting period*), followed by 4 s of grating stimuli (ON period). 8 directions were presented from 0 to 315° to the horizontal, while each direction was displayed for 0.5 s and started with another direction of 45° increment (Figure 5A3).

In vivo experiments: Data processing and analysis

The recorded images stacks were processed in Fiji (ImageJ version 1.53c) before data extraction. The slices from contralateral (*contra*) and ipsilateral (*ipsi*) recordings (with respect to the visual stimuli) were combined and stacked by maximum intensity Z-projection to concatenate into a single movie. The motion artifact was minimal with plugin 'Template Matching' by recognizing and aligning blood vessels over large region within slices. The Z-projected slice was kept for alignment only. The aligned movies were imported to Suite2P[52] (version 0.10.1, https://github.com/MouseLand/suite2p) for neuron segmentation. The following parameters were changed from the default: tau of 0.75, denoise of 1, diameter of 9, anatomical only of 1, maximum iterations of 1, frames per second of 7.5 and 191 minimum neuropil pixels. The regions of interest (ROIs) and corresponding cells' activities were detected and saved in.mat file for further processing in MATLAB.

The ROIs of cell were identified with Suite2P-generated file, *iscell*, and the non-cell components were removed. Z score and the relative change in fluorescent ($\Delta F/F_0$, where F_0 was the fluorescence baseline of that ROI) was calculated. A two-step approach was used to further select the visually responsive neuron, which showed significant activities to specific direction(s). First, the $\Delta F/F_0$ of each direction was averaged within trial and compared to the baseline level averaged over two consecutive seconds (e.g in Figures 5B and 5C, from the start of fourth sec. to end of fifth sec.) using two-sided t-test at 5% significance level. Afterward, among the neurons with significant difference at any direction, those with Z score >3 for at least 3 consecutive frames of ON period were identified. These neurons were regarded as visually responsive. Only these visually responsive neurons were tested with the spike inference algorithm. Before inputting into the ENS^2 system, the $\Delta F/F_0$ signals were processed following the same data preparation procedures described above (i.e. re-sampled to 60Hz and segmented into 96-sized fragments).

In vivo experiments: Calculation of tuning curves and selectivity indexes

For each trial in the recording of responsive neurons, mean responses within each 0.5s stimulus window are taken (e.g. mean $\Delta F/F_0$ for calcium traces and mean firing rate for spike-events predictions), producing 8 different mean response values. We also compute the background responses using the averaged activities within the 6s resting period. Then, the resultant 8 mean response values are subtracted from their corresponding background responses in each trial. The final tuning curves are obtained by averaging these mean responses across 20 trials.

We adopted the OSI (orientation selectivity index) and DSI (direction selectivity index) as defined in a previous study[54] to quantify the tuning curves and neuronal selectivity further,

$$OSI = \left| \frac{\sum_k R(\theta_k) e^{2i\theta_k}}{\sum_k R(\theta_k)} \right| \quad \text{(Equation 11)}$$

$$DSI = \left| \frac{\sum_k R(\theta_k) e^{i\theta_k}}{\sum_k R(\theta_k)} \right| \quad \text{(Equation 12)}$$

where $R(\theta_k)$ is the response to stimulus orientation at angle θ_k, and $k = 8$ is the total number of stimulus angles. Before computing OSI/DSI, if any value in a tuning curve is below zero, we upshift the whole tuning curve to keep it non-negative for calculation of OSI and DSI. The OSI/DSI of the neuron population were quantified with bar plots showing their medians and 95% confidence intervals, and statistically compared (Figure 5E).

Training and validation of Noise-Augmented ENS^2

The Noise-Augmented ENS^2 is trained using the benchmark database. During training, random white-noise is added to each sample in a batch, resulting in noise-level of up to 15. The noise-levels ranging from 0 to 15 (0 means raw input) are randomly selected and assigned to each sample for each batch. No noise is added if the noise-level of the original recording is higher than the assigned one (e.g. 0). Other training criteria remain the same as the original ENS^2 in benchmark.

The white/pink/red noises used for the Noise-Augmented ENS^2 are generated with an open-source toolkits[71] under GNU LGPL license. Time series of noises are freshly generated for each recording, according to the duration and sampling rates, and then imposed on the latter. We control the variances of generated noise series to obtain noisy recordings with designated noise-levels (e.g. 4).

QUANTIFICATION AND STATISTICAL ANALYSIS

During our benchmarks, we recorded the metrics as described in the section of ''Evaluation of spike inference algorithms'' for each neuron in the testing dataset. To compare among different system configurations and models, we show their medians with 95% confidence intervals (Figures 2A–2D, 3A–3D, 4F–4H, 5E, 6N–6Q, 7G, S1A–S1F, S2A–S2D, S3, and S5) among all neurons from all testing datasets. We performed Shapiro-Wilk test before subsequent statistical analyses, and all did not pass normality tests. We then used Friedman test with Dunn's multiple comparison for statistical analyses when the number of paired groups are larger than two (Figures 2A–2D, 3A–3D, 4G, 5E, S2A–S2D, and S3), and two-sided Wilcoxon signed-rank test otherwise (Figures 4F, 4H, 6J–6M, S2J, S2K, and S5). When testing on the benchmark dataset (Figures 2A–2D, 3A–3D, 4F–4H, 6N–6Q, 7G, S1A–S1F, S2A–S2D, S3, and S5), the performance resulted from two competing models on the same testing neuron is regarded as a paired sample. For the visual stimulating experiment (Figure 5E), the selectivity indexes obtained from $\Delta F/F_0$, spike-rate, or spike-event on a same neuron are regarded as a paired sample. We reported the significance with p values (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$, ns: not significant).