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, *], Chung Tin [1, *]

[1] Department of Biomedical Engineering, City University of Hong Kong, Hong Kong SAR.
[2] School of Biomedical Sciences, The Chinese University of Hong Kong, Hong Kong SAR.
[3] Department of Brain and Cognitive Sciences, Picower Institute for Learning and Memory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139 USA.

[*] Corresponding authors: chungtin@cityu.edu.hk, jacqueip@cuhk.edu.hk

ABSTRACT

Calcium imaging technique provides the advantages in monitoring large population 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 raw calcium inputs based on U-Net deep neural network. When testing on a large, ground truth public database, it consistently outperformed state-of-the-arts algorithms in both spike-rate and spike-event predictions with reduced computational load. We further demonstrated that ENS^2 would improve analyses of orientation selectivity in primary visual cortex neurons. We concluded that optimizing our system for spike-event prediction would produce a versatile inference system that benefits diverse neuroscience studies.

Keywords: calcium imaging; spike inference; deep learning; neural networks

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 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 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 non-linear changes. These fluctuations in calcium are again non-linearly 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 light conditions. Furthermore, the high spatial resolution of calcium imaging trades off with indicators' slow temporal dynamics up to hundreds of milliseconds [14,22], resulting 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 addictive 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 Maximize 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 become rather computational expensive to use. On the other hand, data-driven systems based on supervised learning have emerged with promising performance. A supervised deep learning algorithm called CASCADE has been reported recently, which delivered top-ranking spike inference performance when training data with matched noise levels as the testing dataset were selected for training [38]. Previous data-driven models have limited generalization ability due to 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 with the development of CASCADE. It has facilitated such data-driven approach 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 un-seen recordings with high performance is necessary. 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 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 networks on the spike inference tasks. The preferred configurations of input types, network architectures, and cost functions were investigated. Here, we developed the ENS^2 (effective and efficient neural networks for spike inference from calcium signals) system (Fig. 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 primary visual cortex and showed how the spike inference could improve the analyses of the experimental data. Lastly, 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 calcium data that will favor future algorithm development, which could help us understand the complex process in the brain.

RESULTS

Raw calcium input with MSE loss is adequate for spike inferring tasks

For selecting the best method to infer spikes, we evaluated all potential configurations of models across all 20 datasets (see Methods), where they are expected to generalize to un-seen datasets for achieving high performance. First, we tested how differently pre-processed calcium inputs would benefit the inference task. When measured in Corr (correlation, Fig. 2A), raw inputs hold clear leads over either normalized inputs or de-noised inputs. However, no clear advantage was observed when using the other evaluation metrics (Fig. 2B-D & Extended Data Fig. 1A-B).

Next, we assessed if the choice of the loss functions would affect the performance of our models. Here, we compared the case with MSE (mean square error), vRD (van Rossum distance) and Corr as loss functions. Intuitively, we expected that using vRD loss and Corr loss would benefit the performance evaluation using vRD and Corr respectively. Nevertheless, our results show that this is only true for some cases. In fact, the benefit with Corr loss on Corr evaluation is not statistically significant (Fig. 2A). On the other hand, using Corr loss results in undesired vRD, VPD, and ER with either U-Net or FC-Net (Fig. 2B-D), and in fact,

the performance (in terms of vRD, VPD, and ER) is always worse than using vRD loss or MSE loss. Similar observation is found when evaluated with Error and Bias (Extended Data Fig. 1A-B). The major reason of this is that the Corr loss function fails to differentiate predictions of similar temporal pattern but different amplitude (Fig. 2H), in such a way that it tends to make a prediction with minimal spike-rate (in amplitude) as compared to other configurations (Fig. 2F-I). As a consequence, spike-event could not be reliably estimated from the predicted spike-rate due to low signal-to-noise ratio. On the other hand, using vRD loss does not show clear advantage over using MSE loss in terms of performance, even when evaluated with vRD (Fig. 2B). In fact, vRD loss is more costly to use than MSE loss since its computation requires ground truth spike-event in addition to ground truth spike-rate. As a result, we took raw calcium input with MSE loss function in our proposed ENS^2.

U-Net achieves the best overall performance in spike inferring tasks

We then compared three architectures of neural networks (U-Net, Le-Net, and FC-Net, see Methods). As shown in Fig. 2, U-Net delivers the best overall performance, especially for Corr evaluation. When measured by vRD and ER, slight advantages with U-Net could also be observed over the other two networks, in particular with MSE loss or vRD loss. This also holds when we evaluated the models with Error and Bias (Extended Data Fig. 1A-B).

To show that the difference in performance came from the network architecture rather than specific hyper-parameter settings, we repeated the simulation using the configuration of U-Net, raw input, and MSE loss with various hyper-parameters (Extended Data Fig. 2). The filled bars represent the default hyper-parameter combination used in this study as described in the Method section. Regardless, we showed that they all have little effect on the final performance. The Corr approaches 70% and vRD remains less than 3 in all cases. The VPD and ER are around 0.6 and 50%, respectively.

We also proved that our models were trained adequately with our early-stopping criteria (see Method). Extended Data Fig. 2G illustrates the MSE training losses for all 20 datasets. The losses decrease with more iterations generally and are stabilized sufficiently as the training stops. Moreover, Extended Data Fig. 2 demonstrates that the patience of iterations (see Method) before early-stopping has little influence on performance. Together, we proved that our models had been trained and regularized sufficiently by iterating over only thousands of batches of data to avoid over-fitting.

Comparison to state-of-the-art studies

Based on our investigations above, we selected the configuration of U-Net, raw input, and MSE loss as our proposed ENS^2 system. We took this further to compare it with two representative state-of-the-art studies: CASCADE [38] and MLspike [29]. We selected these two methods as they are 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. Results are summarized in Fig. 3.

We first benchmarked their performance across all 20 datasets (see Methods). When measured by Corr, vRD, VPD, and ER, Fig. 3A-D shows that the data-driven systems (i.e. our proposed ENS^2 and CASCADE) generally perform better than the model-based system (i.e. MLspike). We showed here that data-driven models could perform better even in VPD and ER (up to 10%) than the model-based method, which is specially designed for spike-event prediction. This may originate from the generalization abilities of data-driven methods, while model-based method relies on auto-calibration of parameters. When compared to CASCADE, our systems also showed superior performance for both spike-rate prediction (Fig. 3A-B, Extended Data Fig. 1C-D) and spike-event prediction (Fig. 3C-D). In particular, our ENS^2 shows 10% higher in Corr and 5% lower in ER than CASCADE.

We took a deeper look into these results by pooling all the neurons and examining the distribution of the performance indices (Fig. 3E-H). For spike-rate predictions, we found that ENS^2 reached over 80% in Corr and below 2 in vRD for a notable number of neurons, while much fewer neurons for CASCADE and MLspike (Fig. 3E-F). Similar observations can be found in spike-event inferences (Fig. 3G-H). This proves that our ENS^2 system yielded better performance consistently for most neurons. We also investigated how the performance of each algorithm varied for each single neuron (or dataset) (Fig. 3I-L). The comparison shows that the performance of inference for certain neuron (or dataset), regardless of algorithm, indeed depends significantly on their own properties, such that for neurons that were well/badly predicted by one algorithm were usually predicted (relatively) well/badly by the other algorithms as well (e.g. dataset 8/20).

We highlighted several specific segments of recording from four datasets of different frame rates, calcium indicators and species as examples to further understand how the different models and metrics impact the spike predictions (Fig. 4). For dataset 9 with the GCaMP6f indicator and a high sampling rate up to 160Hz (Fig. 4A), CASCADE tended to output broad spike-rate prediction, and thus resulting in over-estimation with shifted spike-events (shaded in orange). In contrast, MLspike tends to under-estimate such that quite a number of missing spike events were observed. On the other hand, our system (ENS^2) showed better predictions than both these two methods for all evaluation metrics (bold values on the right in Fig. 4A). For dataset 12 under a similar high

frame rate but using the GCaMP6s indicator (Fig. 4B), the calcium signals possess diverse dynamics upon spiking. The prediction task becomes challenging because some spike-events could hardly be identified even by visual inspection. In this case, we find our ENS^2 system predicts spike-rates and spike-events that are closest to the ground truth. Similarly in dataset 13 (Fig. 4C), we found excessive spike-event predictions by both CASCADE and MLspike, probably due to the additive noise in noise matching input or shot noise in raw input. For dataset 6 recorded from zebrafish (Fig. 4D), the frame rate is lower (30Hz). The slow calcium dynamics caused over-estimation in all three methods. However, we found that the ENS^2 system best recovered the temporal firing pattern, predicting three more precise spike-rate windows (shaded in orange) with fewer over-estimated spike-events. Similar advantages of our systems could also be observed in Fig. 2E-F, where CASCADE and MLspike also tend to over-estimate and under-estimate in dataset 11, respectively. Here, we show that our proposed systems could maintain robust inference capability under varying conditions.

Fig. 2-4 also show that pre-processing input data did not improve performance of our models apparently. For instance, when predicted on dataset 11, explicitly canceling noise (e.g. de-noised input, Fig. 2I) or matching noise-levels of input data (e.g. CASCADE, Fig. 2F) did not gain any advantage in the benchmark. They consistently showed worse performance in both spike-rate and spike-event predictions than the ENS^2 system with raw input (Fig. 2F). Instead, we found that using raw inputs for our proposed model achieves the best inferring results for most cases. We suggest that our data-driven neural network could handle a considerable range of noise turbulence implicitly by itself, since they perform equally well or even better with raw inputs. This reduces the workload of input pre-processing and feature engineering that were compulsory in traditional machine learning tasks.

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, Fig. 5). For a specific sampling rate (e.g. 60Hz), series of noise matching models were trained in CASCADE to meet the need of different noise levels. Each of their noise matching model consisted of 5 identical networks for ensemble learning to boost performance. On the other hand, only a single network is required in our ENS^2 to predict data under each sampling rate for various conditions. As a result, the ENS^2 with U-Net requires 20k fewer trainable parameters, and only a maximum of 5.12 million data segments are fed for training, which is fewer than that of CASCADE by two orders (Fig. 5A). In particular, even using an entry-level GPU (Nvidia GTX1650), training of ENS^2 on millions of samples from 19 datasets completed in less than 2.5 minutes on average, and it took just seconds for inference on a testing dataset. These features are not only beneficial to off-the-shelf usage for spike inference, but also enable cost-effective re-training or fine-tuning when more paired datasets are available to improve our model further. Our spike-event estimation algorithm (see Methods) is also more computationally efficient than that used in CASCADE (Fig. 5B-D). When applied on the same spike-rates predictions of all

20 datasets from the ENS^2 system, our greedy algorithm performed faster by one order of magnitude (Fig. 5B), while maintaining similar accuracy (Fig. 5C-D). As such, we show that our ENS^2 demonstrated high performance and good computational efficiency for inferring spikes from calcium data than the state-of-art methods. In fact, while CASCADE took 26sec for spike-event estimation alone, this time is sufficient for the ENS^2 system to complete both spike-rate prediction and spike-event estimation.

Application to information encoding in primary visual cortex

In the above benchmark, we show that ENS^2 accomplished relatively high performance and high efficiency for inference in un-seen recordings. Next, we ask if our system indeed helps in real neuroscience problems as previous models did [5,7,8]. Here, we trained our full ENS^2 system with all 20 datasets available, and then deployed it to unseen calcium imaging data recorded from the primary visual cortex (V1) (Fig. 6). We collected *in vivo* calcium fluorescence images from V1 of mice that were shown to drifting grating stimuli of four unique orientations, that move in two opposing directions (8 directions total) (Fig. 6A, see Methods). Responsive neurons were selected, and their fluorescence signals were processed into calcium traces ($\Delta F/F_0$) for further analyses (Fig. 6B1-2 & 6C1-2, see Methods). We then used our ENS^2 system to predict the spike-rate (Fig. 6B3 & 6C3) and spike-event (Fig. 6B4 & 6C4) accordingly. We compared our analyses with these different inputs and verified if 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 Fig. 6B-C.

We first constructed the tuning curves (see Methods) for each neuron (Fig. 6B5-7 & 6C5-7). It should be clear that the resultant tuning curves are quite different when computed using $\Delta F/F_0$ (Fig. 6B5 & 6C5) and using spike-rate or spike-event (Fig. 6B6-7 & 6C6-7). In particular, the spike-rate/spike-event tuning curves have sharpened preferred orientation for these cells, which are beneficial for calcium data analysis and interpretation. The diffused tuning curve by $\Delta F/F_0$ is mainly due to the long "tail" in $\Delta F/F_0$ signal after each peak (Fig. 6B1 & 6C1) due to the slow dynamics of calcium indicators. As a result, there is a shift in the preferred orientation and leads to a misinterpretation of the neuronal behavior. On the other hand, by predicting the spike-rate/spike-event with our ENS^2 system, we reliably eliminated 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 Methods). The neuron in Fig. 6B is a sample cell expected to have high OSI but low DSI; while the neuron in Fig. 6C is expected to have high OSI and high DSI [46] (see Methods & Extended Data Fig. 7). Our results show that OSI computed from $\Delta F/F_0$ was low for both neurons in Fig. 6B-C but was higher when computed from spike-rate/spike event (from 0.47 to 0.77/0.83 and from 0.26 to 0.76/0.71, respectively). Similarly, DSI computed from $\Delta F/F_0$ was low for the neuron in Fig.

6C, but increased when computed from spike-rate/spike event (from 0.36 to 0.62/0.54). Extended Data Fig. 3A-B show two neurons expected to have a preferred orientation at 0°, but with a strong baseline drift in their $\Delta F/F_0$. Due to the shift in baseline, the OSI or DSI tends to be higher when using $\Delta F/F_0$. In contrast, the predicted spikes were able to filter out the shift in baseline, which resulted in a more accurate readout of the OSI/DSI measurement (such as Extended Data Fig. 3B). In all these cases, we verified that the OSI and DSI computed from spike-rate/spike event predicted from our ENS^2 system can better discriminate the response pattern of these neurons than from the raw $\Delta F/F_0$.

In fact, this holds true in general for the neuron population we have recorded (192 contra and 141 ipsi neurons (see Methods)). Fig. 6D1-2 show that OSI computed with spike-event was higher than that computed with $\Delta F/F_0$ in >80% of the cells (89.06% for contra neurons and 82.98% for ipsi neurons). Also, Fig. 6D3-4 show that DSI computed with spike-event was higher than that computed with $\Delta F/F_0$ in >68% of the cells (68.23% for contra neurons and 68.09% for ipsi neurons). Overall, the median OSI/DSI for both contra and ipsi neurons are higher with spike-event/spike-rate than $\Delta F/F_0$ (Fig. 6E). The higher OSI/DSI can increase the sensitivity to discriminate the response pattern of these neurons for their orientation preference. The wider range of OSI/DSI distribution can also increase the dynamic range for analyzing these neurons' characteristics. As such, the improved analytical power using the predicted spike-event/spike-rate from our ENS^2 could benefit the understanding of the complex information processing in the visual cortex.

It is also worth noting how the spike inference could benefit analyses of neurons that have weak responses or where the signal-to-noise ratio (SNR) is low (Extended Data Fig. 3C-D). In Extended Data Fig. 3C, the $\Delta F/F_0$ of this neuron only varied over a range of 0.2 such that the SNR is very low. The resultant noisy tuning curves suggested no orientation or direction selectivity for this neuron. Instead, the spike inference has increased the data SNR such that the tuning curve is much sharpened, showing preferred orientation at 0° and 180° and hence an increased OSI. On the contrary, in Extended Data Fig. 3D, 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 raw $\Delta F/F_0$ signal were filtered out, resulting in negligible OSI and DSI. This suggests 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 has increased the SNR of the recording to not only improving the sensitivity in detecting the orientation selectivity of the neurons, but also to screen out some marginally responsive neurons.

Neuropil correction is a common pre-processing step for improving SNR of calcium imaging signal [12,14,47]. We conducted additional simulations to examine how this pre-processing step blends with the spike inference in analyzing these data. Extended

Data Fig. 4A shows that neuropil correction increased OSI and DSI computed from $\Delta F/F_0$ in more neurons than that computed from predicted spike-event. After neuropil correction, when spike inference with our ENS^2 was performed, more than 60% of the neurons had larger OSI and more than 45% of the neurons had larger DSI than when $\Delta F/F_0$ was used (Extended Data Fig. 4B). Consequently, the median OSI with spike-rate/spike-event were again higher than that with $\Delta F/F_0$, while the three median DSIs were similar in this case (Extended Data Fig. 4C). Nevertheless, the median OSI and DSI with spike-rate/spike-event showed a minor increase between before (Fig. 6E) and after (Extended Data Fig. 4C) neuropil correction. These results suggest that the spike inference with our ENS^2 is by itself effective enough to improve the SNR in the recorded data for better analyses of orientation selectivity in these neurons. As such, additional pre-processing for neuropil correction may be unnecessary.

Factors affecting inference performance

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

Fig. 7A-D show the performance with individual dataset achieved by different configurations of models (the configurations are numbered horizontally in the same order as in Fig. 2A-D). The results show that performance indeed depends strongly on the dataset. For instance, regardless of the networks used, dataset 19 and 20 achieved notably poor vRD than other datasets (Fig. 7B), and some datasets (e.g., 2-3, 14-15, and 19-20) showed considerably worse ER than the others (Fig. 7D). These were also observed when comparing our ENS^2 to state-of-the-art methods (Fig. 3I-L). We extracted various quantities from each dataset, such as noise level, frame rate, firing rate and AP amplitude, and examined how they may affect the inference performance (Fig. 7E-H, Extended Data Fig. 1M-N). It is shown that AP amplitude of the dataset is the key predictor for the inference performance (including Corr, ER, Error, and Bias). We also noticed that the raw AP amplitudes are of critical importance for inference, which relate the number of spike events for a certain calcium indicator. For example, normalization of calcium inputs might be beneficial when calcium signals fluctuate significantly (e.g. Fig. 2E), but would also cause under-estimation or over-estimation since the original AP amplitudes are altered. On the other hand, de-noising the inputs may improve the signal-to-noise (SNR) ratio. But it will also smooth out the calcium traces, causing broader spike-rate outputs and less precise spike-event predictions. Matching the inputs' noise-level with the testing dataset is another potentially useful approach to improve inference performance, but it is not always straight-forward to distinguish the AP amplitude and noise amplitude (e.g. Fig. 2E, Fig. 4B-C). A previous model-based study also testified that an accurate estimate of AP amplitude improved performance [48]. The AP amplitude indeed strongly depends on the calcium indicators' sensitivity. It is clear that one major bottleneck of inference

algorithm is in the calcium indicators.

While Fig. 7A-H suggests that the sensitivity of calcium indicator is a strong predictor for inference performance, we examined how mixing datasets with different types of calcium indicators, like the one used here, may impact the inference performance. We partitioned the training data according to their calcium indicator type with respect to the testing data. Here, "*All*" is the same as the leave-one-dataset-out setting that we have been using in the other part of this paper. "*Same*" includes only those training datasets with the same calcium indicator as the testing data. "*Different*" includes only those training datasets using distinct calcium indicators from the testing data. Fig. 7I-M and Extended Data Fig. 1I-J show that apparently there is no consistent advantage in using dataset with the same calcium indicators for training. Previous study [38] also reported that clustering the same calcium indicators for training showed no advantage. This may suggest that the neural network could calibrate itself over different calcium dynamics in order to generalize to un-seen testing data to certain extent.

Since our ENS^2 is a data-driven model, we wonder how much training data is needed for achieving good inference performance. We randomly sampled different numbers of segments from the total of over 20 hours of available paired data. Here, we ignored the overlapping in these segments and just assumed that around 68,000 64-sized segments under 60Hz were equivalent to 20 hours of recordings. When supplying all available paired data to the model, the total duration is approximately 20 × 64 hours since the paired data is segmented with a step of 1 (see Method). Not surprisingly, the performance of inference increased with the amount of training data but it converged at a maximal level with roughly 5 hours of paired data (Fig. 7N-Q, Extended Data Fig. 1G-H).

DISCUSSION

In this work, we have developed a high performance inference system (ENS^2) through extensive and empirical research, and showed its usefulness in inferring both spike-rates and spike-events.

We have 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. On the other hand, 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, a recent data-driven model (CASCADE [38]) with state-of-art performance also used a network with convolutional layers. Surprisingly, although inferring spikes from calcium signals is a typical sequence-to-sequence translation task, recurrent neural networks (e.g., LSTM) did not show notable advantages [36]. However, in this work, we

used deep convolutional networks with sequence-to-sequence translation ability (e.g., 1D U-Net), and showed that it excelled the other state-of-art systems. The advantages of U-Net may come from its skip-connecting architecture (see Method), which is beneficial to temporal prediction. Recently, a 3D U-Net based model has also been proposed to improve SNR in calcium images and facilitate calcium signal extraction [49]. On the other hand, we revealed in our results that MSE loss could readily regulate the optimization of such models. While 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 be defective in real world tasks. For example, the inferred spike-rates are illy-scaled in amplitude and are unable to recover spike-events faithfully.

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), and hyper-parameter of evaluation metric (e.g. ER window size). The comparisons are summarized in Fig. 8A-D. Fig. 8A shows that the Corr increased consistently with larger smoothing windows. Similar observations can also be found in several recent studies [35,36,50]. This is because the GT spike-rates convolved from the GT spike-events with larger smoothing windows have smoother and broader patterns, which favors the measure of Corr. Fig. 8F shows the GT spike-rates obtained by convolving the GT spike-events in Fig. 8E with varying smoothing window sizes (25ms to 200ms) and their corresponding predictions. Apparently, the smoother and broader waveform of GT spike-rate (with larger smoothing windows) simplifies the prediction task, and it is easier to obtain a high Corr with such simpler and smoother PD spike-rate waveform. This also holds for the spike-rate evaluation of vRD and Error (Fig. 8B & Extended Data Fig. 1E). However, we argue that such resultant "better" performance (e.g. high Corr) would not guarantee meaningful predictions as reflected in the PD spike-events, since multiple GT spike-events could be merged into a single peak of spike-rate (Fig. 8E-F). Instead, the temporal firing patterns could be better predicted with narrower smoothing windows. On the other hand, spike-event predictions (VPD & ER, Fig. 8C-D & Fig. 8G) generally improve with higher sampling rates. This is quite reasonable as smaller bin sizes allow more precise estimation of spike-events from spike-rate predictions. Moreover, when high sampling rates are used (e.g. 30 or 60Hz), VPD and ER would also reduce 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. Fig. 8F). We also analyzed the effect of ER window sizes on ER evaluation (Fig. 8G). As expected, smaller ER window sizes put more demanding evaluation on the algorithm and hence, results in higher ER, which is similar to the findings reported in a previous study [29].

Given these analyses, we suggest that our ENS^2 system should be trained with a sampling rate of 60Hz with 25ms smoothing windows for practical use (labeled with white dashed boxes in Fig. 8A-D & Extended Data Fig. 1E-F). On one hand, preparing

calcium inputs under 60Hz could reduce information loss (see Method), and adopting the 25ms smoothing window could achieve near optimal spike-event inference (Fig. 8C-D). We decided to optimize our system for spike-event prediction rather than spike-rate for the reasons discussed above. On the other hand, the performance of spike-event prediction starts to saturate with this scheme. Further increase in the sampling rate would cause computational over-head while to reduce further the smoothing window size might be harmful to training neural networks with gradient descent. We also repeated our benchmark using Causal smoothing kernels as in CASCADE [38]. Nevertheless, similar performance was obtained in our system (Extended Data Fig. 5). It is worth noting that the CASCADE algorithm [38] was indeed benchmarked under 7.5Hz with 200ms smoothing windows. We also trained our ENS^2 under such conditions (labeled in red dashed boxes in Fig. 8A-D & Extended Data Fig. 1E-F), and it shows that our ENS^2 consistently outperformed the CASCADE algorithm in this way for both spike-rate and spike-event predictions (Extended Data Fig. 6). These results support that our ENS^2 is a versatile and highly effective algorithm for spike inference from calcium signals.

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 primary visual cortex (Fig. 6, Extended Data Fig. 3 & Extended Data Fig. 4). Our results demonstrate that our algorithm has reliably filled the gap between calcium imaging and traditional electrophysiological recording, such that we can perform analyses with both high throughput (from calcium imaging) and high precision (from electrophysiology) in the study of our brain.

METHODS

Benchmark database

In this study, we used the publicly available datasets containing both calcium imaging signals and simultaneously recorded electrophysiological signals from excitatory neurons [13-17,35,38,51-53]. For benchmarking and algorithm development purposes, they were recently compiled by [38] into an extensive database with 21 datasets. Specifically, we adopted dataset #2 to #21 following [38] for a fair comparison, and they are labeled as dataset 1 to 20 in this study as shown in Extended Data Table 1. These 20 datasets cover eight different kinds of calcium indicators, a wide range of frame rates (7.7Hz to 500Hz), and various firing rates (0.2Hz to 5.8Hz on average). Over 20 hours of paired ground truth data (calcium signals and spike-events) were recorded from a total of 230 neurons of either mouse or zebrafish brains.

In each dataset, raw calcium signals are provided as the percentage changes of fluorescence amplitude against baseline ($\Delta F/F_0$),

while individual timestamps label spike-events. We also computed the noise-levels as defined in [38] and listed them in Extended Data Table 1. Furthermore, we presented the increase in $\Delta F/F_0$ induced by one action potential (AP amplitude) for each dataset. The AP amplitude is computed using the averaged calcium kernel, which was extracted from paired ground truth data using the deconvolution function with regularized filter in MATLAB.

Data preparation

1) 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 (Extended Data Table 1), 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 impact of sampling rates on inference results is discussed in this work.

2) Pre-processing of calcium signals

Aside from the raw calcium inputs (where only re-sampling is performed), we also considered several pre-processed inputs. First, since normalization is beneficial in back-propagation [54], we prepared the "normalized inputs" by rescaling the amplitude of the raw inputs to $[0,1]$ on a record-by-record basis. Second, considering that calcium signals are intrinsically noisy [21], we computed the "de-noised inputs" by down-sampling the raw inputs to 6Hz and then up-sampling them to the required sampling rates. This should be sufficient to preserve the spiking properties as we noted that the firing rates of all datasets are well below 6Hz (Extended Data Table 1). Moreover, the "noise matching inputs" were obtained from CASCADE [38] to reproduce results with the CASCADE algorithm, where the algorithm is designed to have the noise-levels of training calcium data matching with those of the testing data.

3) Pre-processing of spike-rate

For a pre-defined sampling rate (e.g. 60Hz), raw timestamps 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. Note that the different pre-processed calcium inputs (raw/normalized/de-noised/noise matching) considered here share identical spike count sequences. 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.

4) Data segmentation

To train the neural networks properly, paired sequences of calcium signals and spike-rates were segmented with a moving step of 1 (Fig. 1A). The length of each segment was set to 64 samples, ensuring that each contains at least 1sec of information (e.g. for a sampling rate of 60Hz, 1sec of data consists of 60 data points). In the case of a sampling rate of 60Hz, a total of >4 million segments of paired data were obtained for training.

Network architectures

We tested three different architectures of neural networks to evaluate their effectiveness in the spike inference task. All of them (U-Net [55], Le-Net [56], and FC-Net (fully-connected network)) are typical representations of their own categories (deep neural network (DNN), convolutional neural network (CNN), and typical neural network (NN)), respectively. We adopted these existing models with minor modifications for 1-D calcium signal inputs. The network architectures are summarized in Extended Data Table 2.

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 [55]. 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 [57] for regularization, since calcium signals with various dynamics may co-exist in a same batch of data. We observed that this regularization helped in model convergence. The Le-Net consists of three convolution layers with kernel sizes of 3, 3, and 6, respectively. Average pooling layers with 2-sized kernels are applied between the three convolution layers. A dropout layer [58] 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 64-sized calcium signal inputs, and output 64-sized spike-rate vectors in a sequence-to-sequence translation manner. During prediction, given that the input data are segmented with steps of 1, each time point is indeed predicted for up to 64 times separately from its adjacent segments. We, thus, were able to average these predictions for a robust final spike-rate output of each time point (Fig. 1A). Spike-event output could 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 (Fig. 5A). They are all randomly initiated to have zero means and standard deviations of 0.02. Leaky ReLU (rectified linear units) with slopes of 0.2 is used as activation functions for all layers except for the output layers, where ReLU is used for non-negative spike-rate prediction.

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. In addition, we used two of the evaluation metrics (see below), correlation (Corr) and van Rossum distance (vRD) [59], as loss functions. Through gradient descents, the models are expected to maximize the performance measured by these two metrics.

The Adam optimizer [60] with a default learning rate of 1e-3 is used for all models. Each model is allowed to update for a maximum of 5000 iterations. In each iteration, a batch of 1024 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 2500 iterations. The resultant models are then ready for prediction. Other details of hyper-parameters and operational environment are summarized in Extended Data Table 3.

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 straight-forward (Fig. 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 and convolve it into a spike-rate vector (*est_rate*) in a same way as we have described above. Then the MSE between the resultant *est_rate* and *pd_rate* is calculated. This step is implemented parallelly for 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 timestamp 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 timestamp 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 in proportional to the duration of recordings. We have validated this spike-events estimation algorithm with the Monte-Carlo importance sampling based algorithm proposed in CASCADE [38].

Evaluation metrics

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:

$$Corr_{GT,PD} = \frac{E[(GT-\mu_{GT})(PD-\mu_{PD})]}{\sigma_{GT}\sigma_{PD}} \tag{1}$$

where GT and PD stand for ground truth and prediction, respectively. Secondly, we use the van Rossum distance (vRD) [59] for the evaluation of spike rates prediction:

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

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

$$Error = \frac{\int |PD - GT| dt}{\int GT dt} \tag{3}$$

$$Bias = \frac{\int |PD - GT|\, dt\, |_{PD(t) > GT(t)} - \int |PD - GT|\, dt\, |_{PD(t) < GT(t)}}{\int GT dt} \tag{4}$$

On the other hand, for measuring spike-event prediction, we adopt the Victor-Purpura distance (VPD) [61]. 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 [29,37,48], which measures the F_1 score of the predicted spike-events:

$$ER_{GT,PD} = 1 - F_1 = 1 - 2\frac{sensitivity \times precision}{sensitivity + precision} \tag{5}$$

$$sensitivity = \frac{true\ positive}{true\ positive + false\ negative} \tag{6}$$

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{7}$$

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.

Implementation of algorithms

As described above, this study involves configurations from three types of calcium inputs, three neural network models, and three loss functions, resulting in a total of 27 configurations of models. The best performing system as evaluated by the six metrics described in the Method section was then benchmarked against the state-of-the-art algorithms, including the data-driven method, CASCADE [38] and the model-based method, MLspike [29].

Our simulations followed the leave-one-dataset-out protocol. In brief, the model is first trained on 19 datasets and tested on the remaining one. This is repeated 20 times such that all the 20 datasets are tested respectively. We then recorded the dataset-wise performance and the neuron-wise performance in each testing dataset.

For the CASCADE algorithm, we follow the protocol as described [38]. For each dataset, noise matching inputs are constructed (artificial noise is added to the 19 training datasets to match the noise-level of the remaining one tested on), and five identical models are trained separately for 10 epochs. The averaged outputs of these five models are regarded as the final spike-rate predictions. Spike-event predictions are estimated using a Monte-Carlo importance sampling based algorithm in CASCADE [38].

For the MLspike algorithm [29], raw calcium inputs are used. The drift ($\eta = 0.01$) and non-linearity parameters (saturation $\gamma = 0.1$ for synthetic dyes, polynomial coefficient $[p_2, p_3] = [1.0,0.0]$ for the rest indicators) are set manually with prior knowledge. The values of the model parameters of A (action potential (AP) amplitude), τ (calcium decay time constant), and σ (noise amplitude) are obtained using its built-in auto-calibration algorithm, since manual calibration is impossible in actual usage without ground truth paired data. For a fair comparison, in addition to the direct spike-event outputs from MLspike, we also convolve them into spike-rate outputs in the same way as described in the Method section.

In vivo experiments

1) 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 anaesthetized 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 (Fig. 6A3) 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.

2) 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. 20x Olympus objective lens was used for functional imaging. The calcium signals of neurons from layer 2/3 were visualized at 920nm with acquisition frame rate around

7.6Hz (averaging from 4 consecutive frames around 30Hz).

3) 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 (Fig. 6A2-3). Each trial of display lasts for 10 seconds and repeated for 20 trials per stimulus, resulting in imaging session of 200 seconds. At the beginning of each trial, 6 seconds of grey screen is presented as blank, followed by 4 seconds of grating stimuli (ON period). 8 directions were presented from 0 to 315 degrees to the horizontal, while each direction was displayed for 0.5 second and started with another direction of 45 degrees increment (Fig. 6A3).

4) Data processing and analysis

The recorded images stacks were processed in Fiji (ImageJ version 1.53c) before data extraction. The slices from contralateral and ipsilateral recordings 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 [62] (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 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 with the corresponding $\Delta F/F_0$ of 4th – 6th second of blank with two-sided t-test at 5% significance level. Afterwards, the neurons with significant difference at any direction were searched for exactly 3 consecutive frames of ON period with z-score > 3. The neuron passed both criteria is regarded as visually responsive.

5) Calculation of tuning curves and selectivity indexes

For each trial in the recording of responsive neurons, mean responses within each 0.5sec 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 6sec resting window. 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 [46] to quantify the tuning curves and neuronal selectivity further,

$$OSI = \left| \frac{\sum_k R(\theta_k) e^{2i\theta_k}}{\sum_k R(\theta_k)} \right| \tag{8}$$

$$DSI = \left| \frac{\sum_k R(\theta_k) e^{i\theta_k}}{\sum_k R(\theta_k)} \right| \tag{9}$$

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 up-shift the whole tuning curve to keep it non-negative for calculation of OSI and DSI. Extended Data Fig. 7 shows several representative examples of how different patterns in tuning curves could affect OSI/DSI quantitatively.

Reference

1 Neher, E. & Sakmann, B. Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature* **260**, 799-802, doi:10.1038/260799a0 (1976).

2 Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. *Pflügers Archiv* **391**, 85-100, doi:10.1007/BF00656997 (1981).

3 Spira, M. E. & Hai, A. Multi-electrode array technologies for neuroscience and cardiology. *Nat. Nanotechnol.* **8**, 83-94, doi:10.1038/nnano.2012.265 (2013).

4 Buzsáki, G. Large-scale recording of neuronal ensembles. *Nat. Neurosci.* **7**, 446-451, doi:10.1038/nn1233 (2004).

5 de Vries, S. E. J. *et al.* A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nat. Neurosci.* **23**, 138-151, doi:10.1038/s41593-019-0550-9 (2020).

6 Kerr, J. N. D. & Denk, W. Imaging in vivo: watching the brain in action. *Nat. Rev. Neurosci.* **9**, 195-205, doi:10.1038/nrn2338 (2008).

7 Wilson, N. R., Runyan, C. A., Wang, F. L. & Sur, M. Division and subtraction by distinct cortical inhibitory networks in vivo. *Nature* **488**, 343-348, doi:10.1038/nature11347 (2012).

8 Rikhye, R. V. & Sur, M. Spatial Correlations in Natural Scenes Modulate Response Reliability in Mouse Visual Cortex. *The Journal of Neuroscience* **35**, 14661, doi:10.1523/JNEUROSCI.1660-15.2015 (2015).

9 Giovannucci, A. *et al.* Cerebellar granule cells acquire a widespread predictive feedback signal during motor learning. *Nat. Neurosci.* **20**, 727, doi:10.1038/nn.4531 (2017).

10 Knogler, L. D., Markov, D. A., Dragomir, E. I., Štih, V. & Portugues, R. Sensorimotor Representations in Cerebellar Granule Cells in Larval Zebrafish Are Dense, Spatially Organized, and Non-temporally Patterned. *Current Biology* **27**, 1288-1302, doi:10.1016/j.cub.2017.03.029 (2017).

11 Wagner, M. J., Kim, T. H., Savall, J., Schnitzer, M. J. & Luo, L. Cerebellar granule cells encode the expectation of reward. *Nature* **544**, 96, doi:10.1038/nature21726 (2017).

12 El-Boustani, S. *et al.* Locally coordinated synaptic plasticity of visual cortex neurons in vivo. *Science* **360**, 1349-1354, doi:10.1126/science.aao0862 (2018).

13 Akerboom, J. *et al.* Optimization of a GCaMP Calcium Indicator for Neural Activity Imaging. *The Journal of Neuroscience* **32**, 13819, doi:10.1523/JNEUROSCI.2601-12.2012 (2012).

14 Chen, T.-W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295-300, doi:10.1038/nature12354 (2013).

15 Tada, M., Takeuchi, A., Hashizume, M., Kitamura, K. & Kano, M. A highly sensitive fluorescent indicator dye for calcium imaging of neural activity in vitro and in vivo. *European Journal of Neuroscience* **39**, 1720-1728, doi:10.1111/ejn.12476 (2014).

16 Dana, H. *et al.* Sensitive red protein calcium indicators for imaging neural activity. *eLife* **5**, e12727, doi:10.7554/eLife.12727 (2016).

17 Bethge, P. *et al.* An R-CaMP1.07 reporter mouse for cell-type-specific expression of a sensitive red fluorescent calcium indicator. *PLoS One* **12**, e0179460, doi:10.1371/journal.pone.0179460 (2017).

18 Denk, W., Strickler, J. H. & Webb, W. W. Two-photon laser scanning fluorescence microscopy. *Science* **248**, 73, doi:10.1126/science.2321027 (1990).

19 Stosiek, C., Garaschuk, O., Holthoff, K. & Konnerth, A. In vivo two-photon calcium imaging of neuronal networks. *Proceedings of the National Academy of Sciences* **100**, 7319, doi:10.1073/pnas.1232232100 (2003).

20 Sofroniew, N. J., Flickinger, D., King, J. & Svoboda, K. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *eLife* **5**, e14472, doi:10.7554/eLife.14472 (2016).

21 Grienberger, C. & Konnerth, A. Imaging Calcium in Neurons. *Neuron* **73**, 862-885, doi:10.1016/j.neuron.2012.02.011 (2012).

22 Kerr, J. N. D., Greenberg, D. & Helmchen, F. Imaging input and output of neocortical networks in vivo. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 14063, doi:10.1073/pnas.0506029102 (2005).

23 Yaksi, E. & Friedrich, R. W. Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca2+ imaging. *Nat. Methods* **3**, 377-383, doi:10.1038/nmeth874 (2006).

24 Greenberg, D. S., Houweling, A. R. & Kerr, J. N. D. Population imaging of ongoing neuronal activity in the visual cortex of awake rats. *Nat. Neurosci.* **11**, 749-751, doi:10.1038/nn.2140 (2008).

25 Grewe, B. F., Langer, D., Kasper, H., Kampa, B. M. & Helmchen, F. High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. *Nat. Methods* **7**, 399-405, doi:10.1038/nmeth.1453 (2010).

26 Vogelstein, J. T. *et al.* Fast Nonnegative Deconvolution for Spike Train Inference From Population Calcium Imaging. *Journal of Neurophysiology* **104**, 3691-3704, doi:10.1152/jn.01073.2009 (2010).

27 Onativia, J., Schultz, S. R. & Dragotti, P. L. A finite rate of innovation algorithm for fast and accurate spike detection from two-photon calcium imaging. *J. Neural Eng.* **10**, 046017 (2013).

28 Pnevmatikakis, E. A., Merel, J., Pakman, A. & Paninski, L. in *2013 Asilomar Conference on Signals, Systems and Computers.* 349-353.

29 Deneux, T. *et al.* Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo. *Nat. Commun.* **7**, 12190, doi:10.1038/ncomms12190 (2016).

30 Friedrich, J. & Paninski, L. Fast active set methods for online spike inference from calcium imaging. *Advances In Neural Information Processing Systems* **29** (2016).

31 Pnevmatikakis, Eftychios A. *et al.* Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data. *Neuron* **89**, 285-299, doi:10.1016/j.neuron.2015.11.037 (2016).

32 Friedrich, J., Zhou, P. & Paninski, L. Fast online deconvolution of calcium imaging data. *PLoS Comput. Biol.* **13**, e1005423, doi:10.1371/journal.pcbi.1005423 (2017).

33 Pachitariu, M., Stringer, C. & Harris, K. D. Robustness of Spike Deconvolution for Neuronal Calcium Imaging. *The Journal of Neuroscience* **38**, 7976, doi:10.1523/JNEUROSCI.3339-17.2018 (2018).

34 Jewell, S. W., Hocking, T. D., Fearnhead, P. & Witten, D. M. Fast nonconvex deconvolution of calcium imaging data. *Biostatistics* **21**, 709-726, doi:10.1093/biostatistics/kxy083 (2020).

35 Theis, L. *et al.* Benchmarking Spike Rate Inference in Population Calcium Imaging. *Neuron* **90**, 471-482, doi:10.1016/j.neuron.2016.04.014 (2016).

36 Berens, P. *et al.* Community-based benchmarking improves spike rate inference from two-photon calcium imaging data. *PLoS Comput. Biol.* **14**, e1006157, doi:10.1371/journal.pcbi.1006157 (2018).

37 Hoang, H. *et al.* Improved hyperacuity estimation of spike timing from calcium imaging. *Sci. Rep.* **10**, 17844, doi:10.1038/s41598-020-74672-y (2020).

38 Rupprecht, P. *et al.* A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging. *Nat. Neurosci.*, doi:10.1038/s41593-021-00895-5 (2021).

39 Sebastian, J., Sur, M., Murthy, H. A. & Magimai-Doss, M. Signal-to-signal neural networks for improved spike estimation from calcium imaging data. *PLoS Comput. Biol.* **17**, e1007921, doi:10.1371/journal.pcbi.1007921 (2021).

40 Tsutsumi, S. *et al.* Structure-function relationships between aldolase C/zebrin II expression and complex spike synchrony in the cerebellum. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **35**, 843-852, doi:10.1523/JNEUROSCI.2170-14.2015 (2015).

41 Sebastian, J., Kumar, M. G., Viraraghavan, V. S., Sur, M. & Murthy, H. A. Spike Estimation From Fluorescence Signals Using High-Resolution Property of Group Delay. *IEEE Trans. Signal Process.* **67**, 2923-2936, doi:10.1109/TSP.2019.2908913 (2019).

42 Zhai, X. & Tin, C. Automated ECG classification using dual heartbeat coupling based on convolutional neural network. *IEEE Access* **6**, 27465-27472, doi:10.1109/access.2018.2833841 (2018).
43 Zhai, X., Zhou, Z. & Tin, C. Semi-supervised learning for ECG classification without patient-specific labeled data. *Expert Syst. Appl.* **158**, 113411, doi:10.1016/j.eswa.2020.113411 (2020).
44 Zhou, Z., Zhai, X. & Tin, C. Fully automatic electrocardiogram classification system based on generative adversarial network with auxiliary classifier. *Expert Syst. Appl.* **174**, 114809, doi:10.1016/j.eswa.2021.114809 (2021).
45 Zhai, X., Jelfs, B., Chan, R. H. M. & Tin, C. Self-recalibrating surface EMG pattern recognition for neuroprosthesis control based on convolutional neural network. *Front. Neurosci.* **11**, doi:10.3389/fnins.2017.00379 (2017).
46 Mazurek, M., Kager, M. & Van Hooser, S. D. Robust quantification of orientation selectivity and direction selectivity. *Front. Neural Circuits* **8**, 92-92, doi:10.3389/fncir.2014.00092 (2014).
47 Kerlin, A. M., Andermann, M. L., Berezovskii, V. K. & Reid, R. C. Broadly Tuned Response Properties of Diverse Inhibitory Neuron Subtypes in Mouse Visual Cortex. *Neuron* **67**, 858-871, doi:10.1016/j.neuron.2010.08.002 (2010).
48 Éltes, T., Szoboszlay, M., Kerti-Szigeti, K. & Nusser, Z. Improved spike inference accuracy by estimating the peak amplitude of unitary [Ca2+] transients in weakly GCaMP6f-expressing hippocampal pyramidal cells. *The Journal of Physiology* **597**, 2925-2947, doi:10.1113/jp277681 (2019).
49 Li, X. *et al.* Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised learning. *bioRxiv*, 2020.2011.2016.383984, doi:10.1101/2020.11.16.383984 (2020).
50 Stringer, C. & Pachitariu, M. Computational processing of neural recordings from calcium imaging data. *Curr. Opin. Neurobiol.* **55**, 22-31, doi:10.1016/j.conb.2018.11.005 (2019).
51 Huang, L. *et al.* Relationship between simultaneously recorded spiking activity and fluorescence signal in GCaMP6 transgenic mice. *bioRxiv*, 788802, doi:10.1101/788802 (2020).
52 Khan, A. G. *et al.* Distinct learning-induced changes in stimulus selectivity and interactions of GABAergic interneuron classes in visual cortex. *Nat. Neurosci.* **21**, 851-859, doi:10.1038/s41593-018-0143-z (2018).
53 Schoenfeld, G., Carta, S., Rupprecht, P., Ayaz, A. & Helmchen, F. In vivo calcium imaging of CA3 pyramidal neuron populations in adult mouse hippocampus. *bioRxiv*, 2021.2001.2021.427642, doi:10.1101/2021.01.21.427642 (2021).
54 Lecun, Y., Bottou, L., Orr, G. B. & MÜLler, K. R. in *NIPS workshop (1996) Lecture notes in computer science* 9-50 (Springer, Berlin, 1998).
55 Ronneberger, O., Fischer, P. & Brox, T. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* (eds Nassir Navab, Joachim Hornegger, William M. Wells, & Alejandro F. Frangi) 234-241 (Springer International Publishing).
56 Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278-2324, doi:10.1109/5.726791 (1998).
57 Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).
58 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**, 1929-1958 (2014).
59 van Rossum, M. C. A novel spike distance. *Neural Comput* **13**, 751-763, doi:10.1162/089976601300014321 (2001).
60 Kingma, D. P. & Ba, J. in *3rd International Conference on Learning Representations (ICLR)* (San Diego, CA, USA, 2015).
61 Victor, J. D. & Purpura, K. P. Nature and precision of temporal coding in visual cortex: a metric-space analysis. *Journal of Neurophysiology* **76**, 1310-1326, doi:10.1152/jn.1996.76.2.1310 (1996).
62 Pachitariu, M. *et al.* Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *bioRxiv*, 061507, doi:10.1101/061507 (2017).

ACKNOWLEDGEMENT

This work was supported by Research Grants Council of Hong Kong SAR Project CityU 11104220 (C.T.), ECS CUHK 24117220 (J.I.), and City University of Hong Kong (Project 7005645), Lo Kwee-Seong Biomedical Research Fund (J.I.) and Faculty Innovation Awards (FIA2020/A/04) from the Faculty of Medicine, CUHK (J.I.). Data and source codes to replicate the primary results will be shared upon reasonable request.

AUTHOR CONTRIBUTIONS

C.T. conceived the study. Z.Z. designed the algorithms and performed analyses. J.I. and K.T. collected the *in vivo* calcium imaging data. J.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 the authors contributed to interpreting the results and writing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

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, *], Chung Tin [1, *]

[1] Department of Biomedical Engineering, City University of Hong Kong, Hong Kong SAR.
[2] School of Biomedical Sciences, The Chinese University of Hong Kong, Hong Kong SAR.
[3] Department of Brain and Cognitive Sciences, Picower Institute for Learning and Memory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139 USA.

[*] Corresponding authors: chungtin@cityu.edu.hk, jacqueip@cuhk.edu.hk

FIGURES AND TABLES

Fig. 1. Schematic workflow of the proposed ENS^2 system.

Fig. 2. Performance comparison under various model architectures, loss functions, and pre-processed inputs.

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

Fig. 4. Examples of spike-rates and spike-events prediction from our proposed ENS^2 and state-of-the-art methods.

Fig. 5. Computational complexity of ENS^2 and the state-of-art algorithm.

Fig. 6. Spike inference with ENS^2 for calcium imaging data collected in primary visual cortex (V1) in a visual stimulating experiment.

Fig. 7. Effect of dataset properties, calcium indicators, and data quantity on spike inference performance.

Fig. 8. Effect of evaluation metrics on spike inference performance.

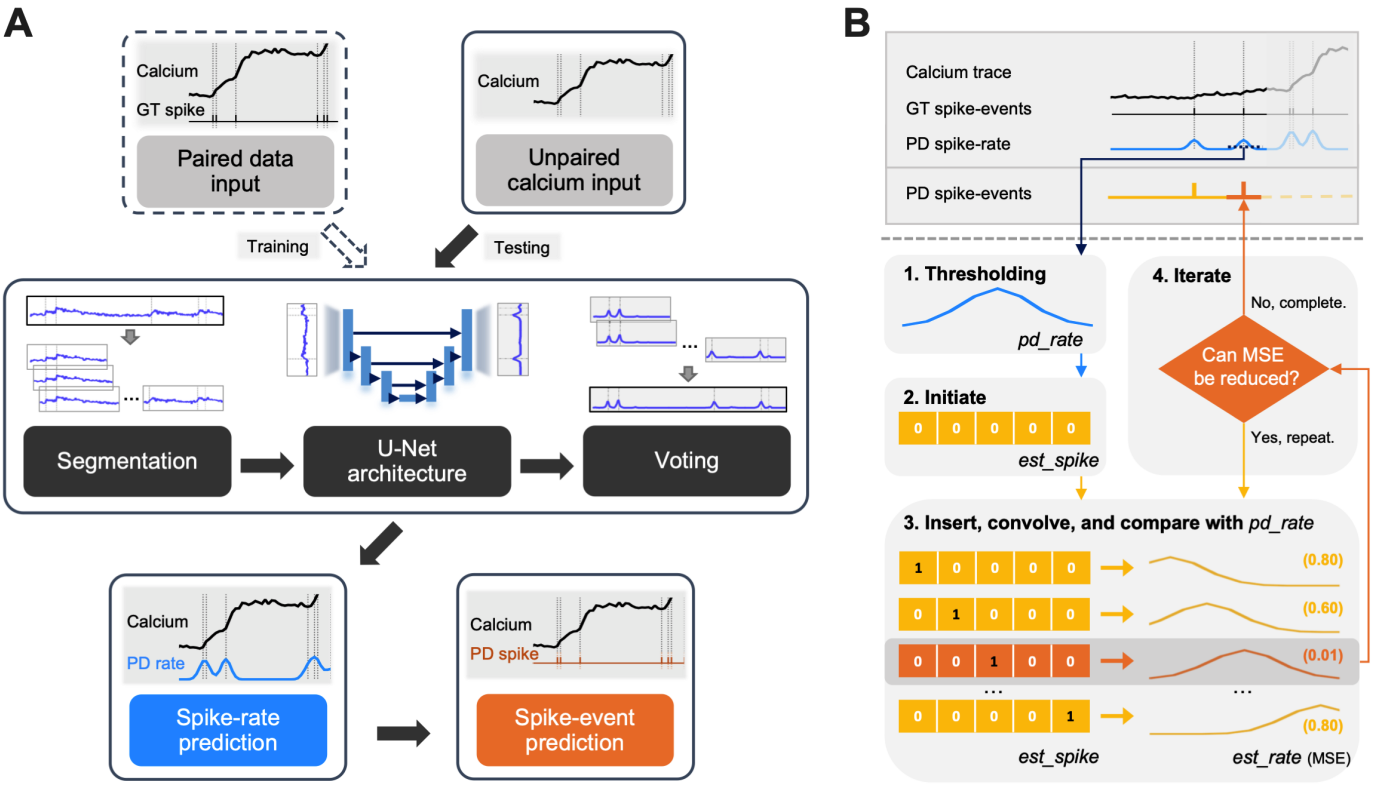

Fig. 1. Schematic workflow of the proposed ENS^2 system. The ENS^2 system contains a neural network to infer spike-rate from calcium inputs (**A**) and an unsupervised greedy algorithm for estimating spike-events from spike-rate predictions (**B**). The neural network is trained with calcium trace inputs paired with ground truth (GT) spike-events, while 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**). Inputs are segmented and fed to the U-Net based model, whose spike-rate predictions are gathered through a voting 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, spike-event is inserted tentatively to approximate the extracted spike-rate fragment. The resultant spike-events sequence that achieves the minimal MSE is regarded as the final spike-events prediction. Details are explained in the Methods section.

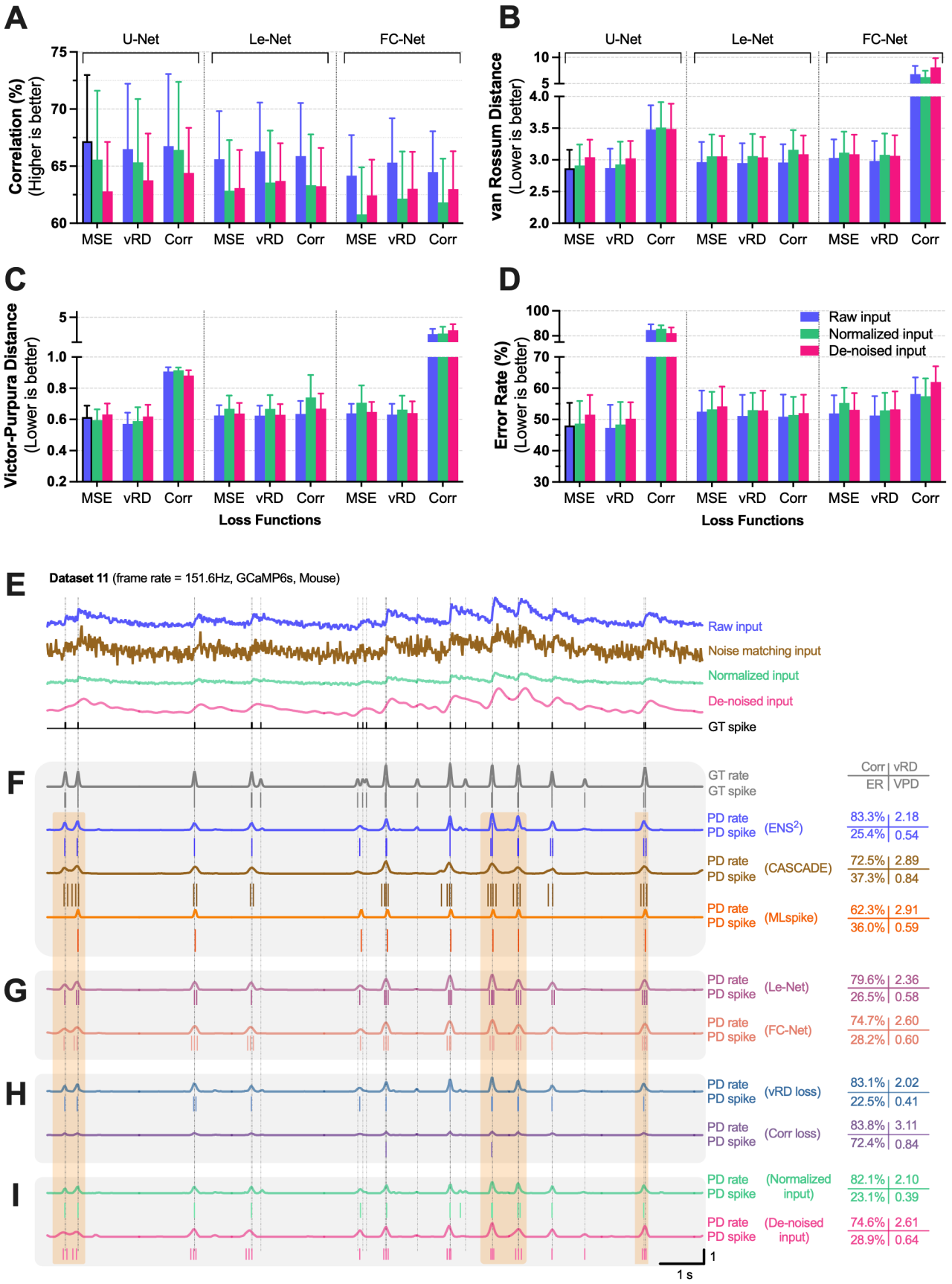
A
U-Net
Le-Net
FC-Net
Correlation (%)
(Higher is better)
MSE
vRD
Corr
B
van Rossum Distance
(Lower is better)
C
Victor-Purpura Distance
(Lower is better)
Loss Functions
D
Error Rate (%)
(Lower is better)
Raw input
Normalized input
De-noised input
E
Dataset 11 (frame rate = 151.6Hz, GCaMP6s, Mouse)
Raw input
Noise matching input
Normalized input
De-noised input
GT spike
F
GT rate
GT spike
Corr | vRD
ER | VPD
PD rate
PD spike
(ENS2)
83.3% | 2.18
25.4% | 0.54
(CASCADE)
72.5% | 2.89
37.3% | 0.84
(MLspike)
62.3% | 2.91
36.0% | 0.59
G
(Le-Net)
79.6% | 2.36
26.5% | 0.58
(FC-Net)
74.7% | 2.60
28.2% | 0.60
H
(vRD loss)
83.1% | 2.02
22.5% | 0.41
(Corr loss)
83.8% | 3.11
72.4% | 0.84
I
(Normalized input)
82.1% | 2.10
23.1% | 0.39
(De-noised input)
74.6% | 2.61
28.9% | 0.64
1 s

Fig. 2. Performance comparison under various model architectures, loss functions, and pre-processed inputs. Performances of our ENS^2 system are measured in (**A**) correlation, (**B**) van Rossum distance, (**C**) Victor-Purpura distance, and (**D**) error rate, respectively. Generally, the configuration of U-Net, MSE loss and raw input provides the best overall performance for both spike-rate prediction and spike-event prediction. (**E**) Examples of different pre-processed calcium inputs (colored) and ground truth spike-events (black). (**F**) Examples of spike-rates and spike-events predicted by our proposed method and state-of-the-art methods. (**G-I**) Examples of spike-rates and spike-events predicted by our proposed method with different configurations. Bar plots present means with 95% confidence intervals. Metrics shown in **F**-**I** measure the performance on the whole dataset. Orange shaded areas represent regions of interest where discrepancies in predictions are observed among different methods.

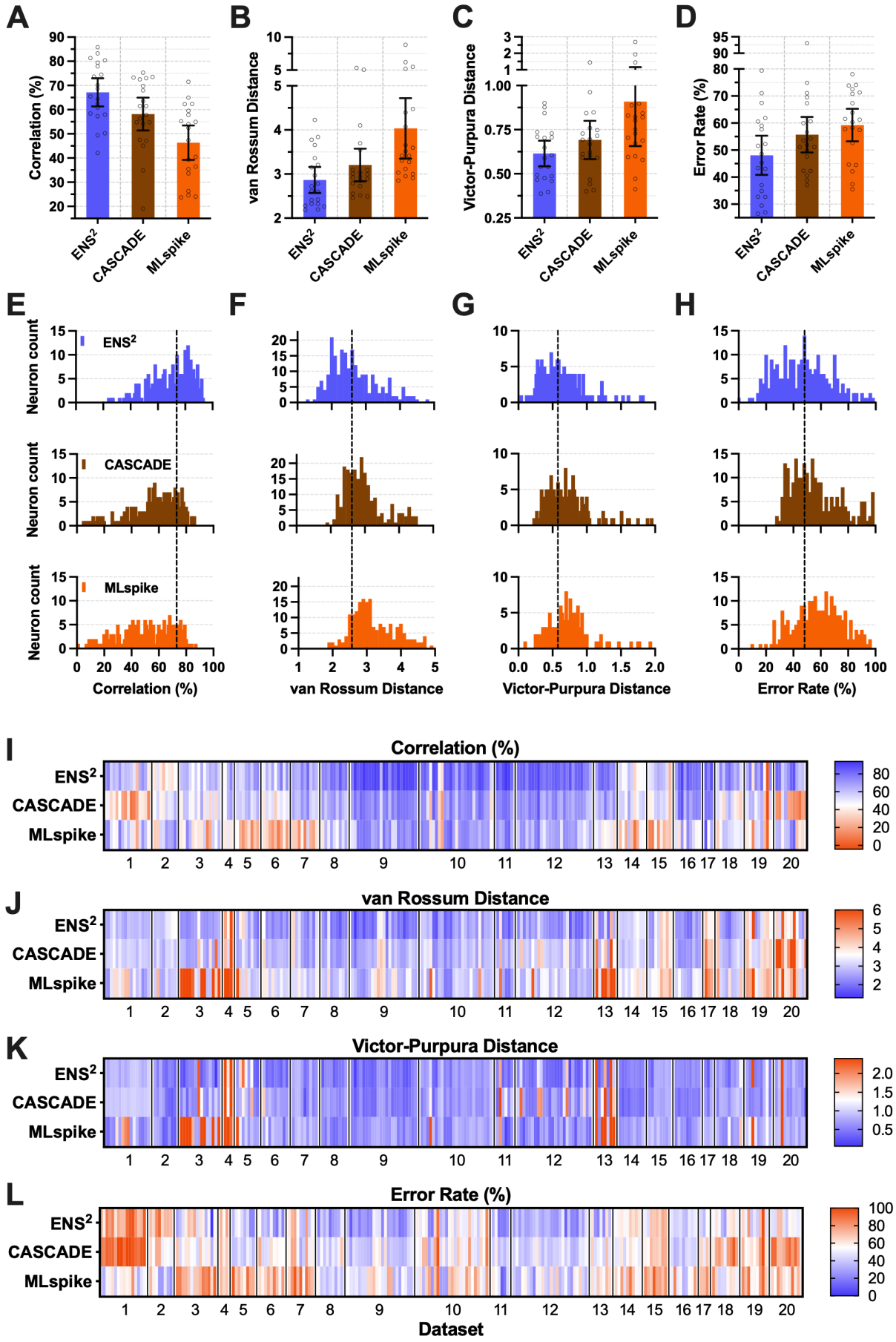
A
Correlation (%)
ENS²
CASCADE
MLspike
B
van Rossum Distance
C
Victor-Purpura Distance
D
Error Rate (%)
E
Neuron count
F
G
H
Correlation (%)
van Rossum Distance
Victor-Purpura Distance
Error Rate (%)
I
Correlation (%)
J
van Rossum Distance
K
Victor-Purpura Distance
L
Error Rate (%)
Dataset

Fig. 3. Performance comparison of our proposed ENS^2 system against state-of-the-art algorithms. Dataset-wise performance is measured in (**A**) correlation, (**B**) van Rossum distance, (**C**) Victor-Purpura distance, and (**D**) error rate, respectively. Inference performance on each neuron is summarized in (**E**-**H**) histograms and (**I**-**L**) heatmap. Empty dots in **A**-**D** denote the performance for each individual dataset. Boxplots show the median performance of the respective algorithms. Vertical dashed lines in **E**-**H** mark the median performance of ENS^2.

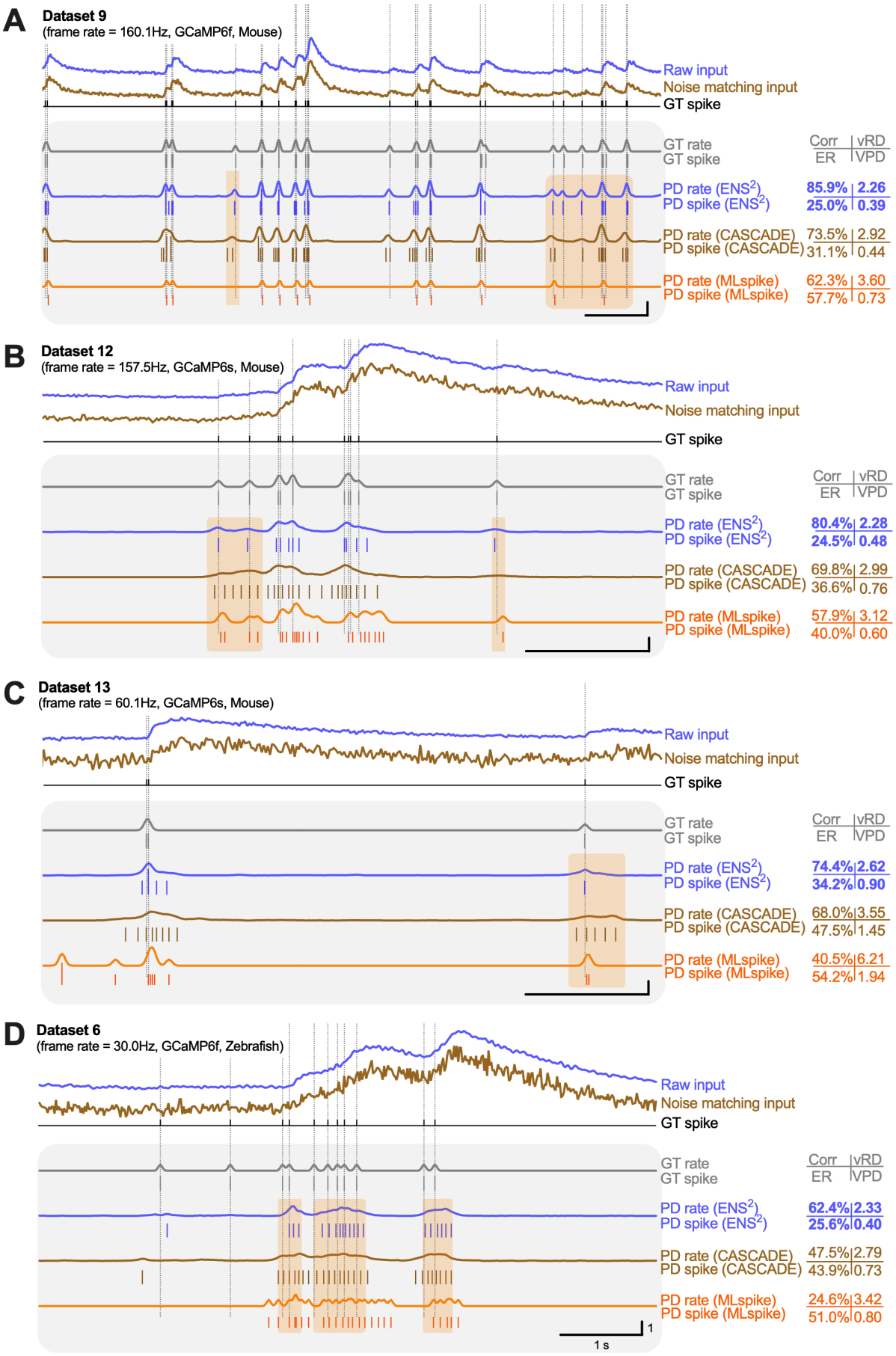
A
Dataset 9
(frame rate = 160.1Hz, GCaMP6f, Mouse)
Raw input
Noise matching input
GT spike
GT rate
GT spike
Corr | vRD
ER | VPD
PD rate (ENS2)
PD spike (ENS2)
85.9% | 2.26
25.0% | 0.39
PD rate (CASCADE)
PD spike (CASCADE)
73.5% | 2.92
31.1% | 0.44
PD rate (MLspike)
PD spike (MLspike)
62.3% | 3.60
57.7% | 0.73
B
Dataset 12
(frame rate = 157.5Hz, GCaMP6s, Mouse)
Raw input
Noise matching input
GT spike
GT rate
GT spike
Corr | vRD
ER | VPD
PD rate (ENS2)
PD spike (ENS2)
80.4% | 2.28
24.5% | 0.48
PD rate (CASCADE)
PD spike (CASCADE)
69.8% | 2.99
36.6% | 0.76
PD rate (MLspike)
PD spike (MLspike)
57.9% | 3.12
40.0% | 0.60
C
Dataset 13
(frame rate = 60.1Hz, GCaMP6s, Mouse)
Raw input
Noise matching input
GT spike
GT rate
GT spike
Corr | vRD
ER | VPD
PD rate (ENS2)
PD spike (ENS2)
74.4% | 2.62
34.2% | 0.90
PD rate (CASCADE)
PD spike (CASCADE)
68.0% | 3.55
47.5% | 1.45
PD rate (MLspike)
PD spike (MLspike)
40.5% | 6.21
54.2% | 1.94
D
Dataset 6
(frame rate = 30.0Hz, GCaMP6f, Zebrafish)
Raw input
Noise matching input
GT spike
GT rate
GT spike
Corr | vRD
ER | VPD
PD rate (ENS2)
PD spike (ENS2)
62.4% | 2.33
25.6% | 0.40
PD rate (CASCADE)
PD spike (CASCADE)
47.5% | 2.79
43.9% | 0.73
PD rate (MLspike)
PD spike (MLspike)
24.6% | 3.42
51.0% | 0.80
1
1 s

Fig. 4. Examples of spike-rates and spike-events prediction from our proposed ENS^2 and state-of-the-art methods. In each subfigure, raw calcium inputs and noise matching inputs are shown with the ground truth spike-events on top. The predicted spike-rates and corresponding spike-events by various methods are shown below. Metrics on the right measure the performance on the whole dataset. Orange shaded areas represent regions of interest where discrepancies in predictions are significant among different methods.

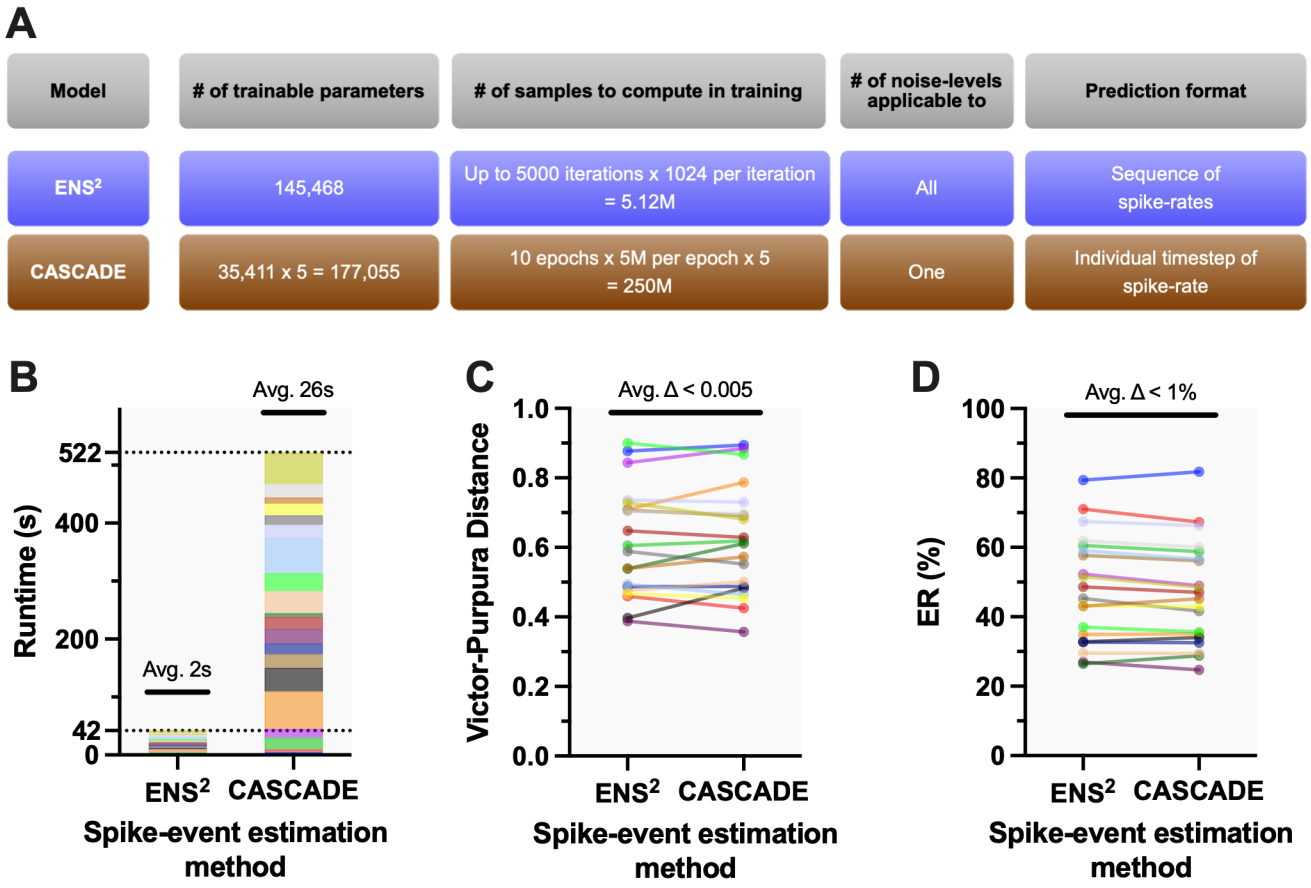

Fig. 5. Computational complexity of ENS^2 and the state-of-art algorithm. (**A**) Comparison of neural networks adopted in ENS^2 and CASCADE. Our method requires two-order fewer samples to compute, and could generalize to all noise-levels. (**B-D**) Comparison of the spike-event estimation algorithm (**Fig. 1B**) in ENS^2 and CASCADE. (**B**) Comparison of runtime spent between the spike-event estimation algorithms in ENS^2 and CASCADE, on estimating spike-events for all 20 datasets under 60Hz. The greedy estimation algorithm in ENS^2 is over one order faster than that in CASCADE. (**C**) VPD measured from spike-events estimated by the two different algorithms, showing comparable estimation performance. Both methods used the same spike-rate predictions from ENS^2 as inputs in this comparison. (**D**) Same as (**C**), but for ER measurement. Results are presented for each dataset. The runtime was measured on a PC with an Intel Core i7-4770 CPU.

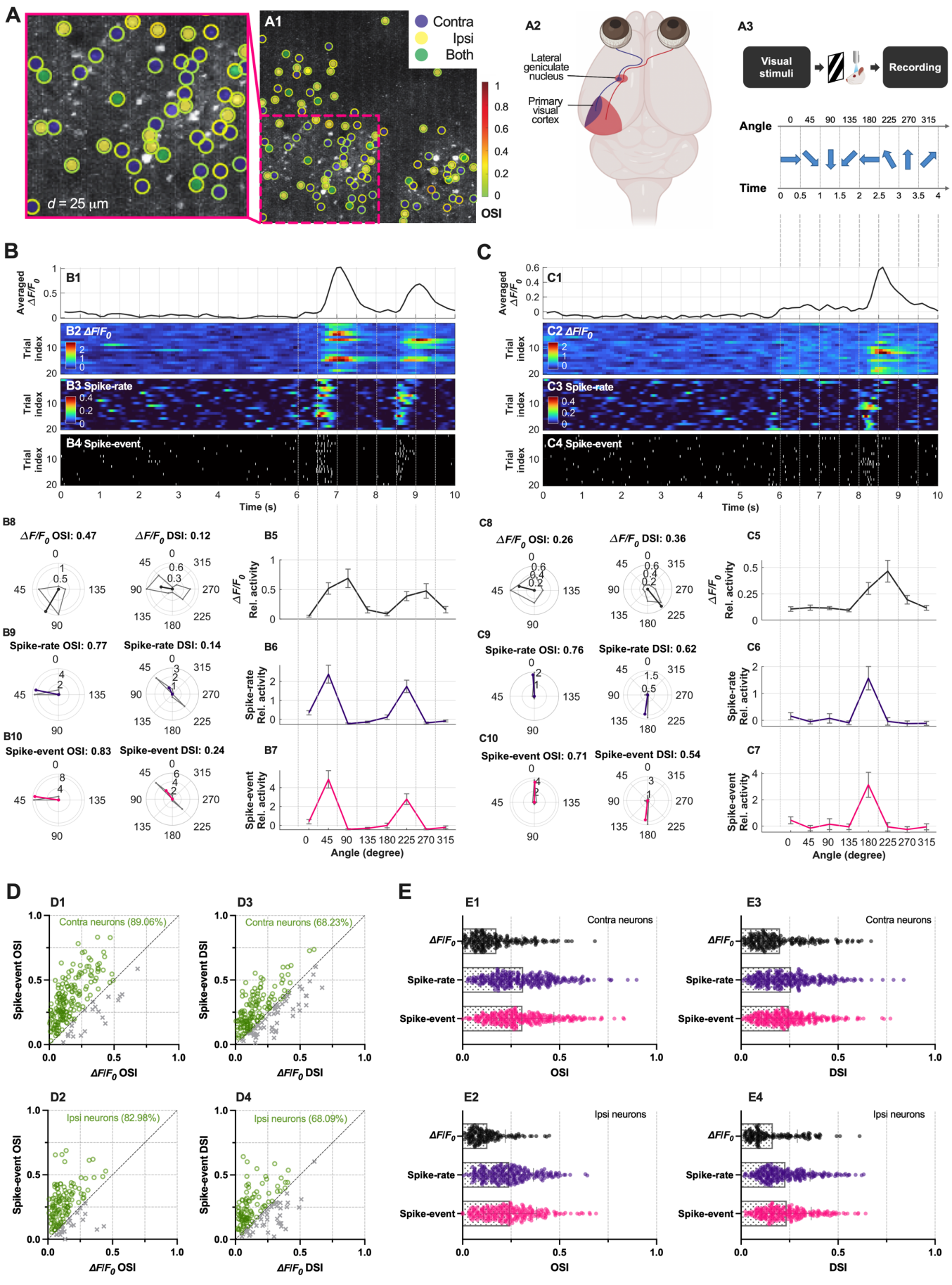
A
A1
Contra
Ipsi
Both
d = 25 µm
OSI
A2
Lateral geniculate nucleus
Primary visual cortex
A3
Visual stimuli
Recording
Angle
Time
B
B1
B2 ΔF/F0
B3 Spike-rate
B4 Spike-event
Averaged ΔF/F0
Trial index
Time (s)
C
C1
C2 ΔF/F0
C3 Spike-rate
C4 Spike-event
B8
ΔF/F0 OSI: 0.47
ΔF/F0 DSI: 0.12
B5
B9
Spike-rate OSI: 0.77
Spike-rate DSI: 0.14
B6
B10
Spike-event OSI: 0.83
Spike-event DSI: 0.24
B7
Rel. activity
Angle (degree)
C8
ΔF/F0 OSI: 0.26
ΔF/F0 DSI: 0.36
C5
C9
Spike-rate OSI: 0.76
Spike-rate DSI: 0.62
C6
C10
Spike-event OSI: 0.71
Spike-event DSI: 0.54
C7
D
D1
Contra neurons (89.06%)
D3
Contra neurons (68.23%)
D2
Ipsi neurons (82.98%)
D4
Ipsi neurons (68.09%)
Spike-event OSI
Spike-event DSI
ΔF/F0 OSI
ΔF/F0 DSI
E
E1
Contra neurons
E3
E2
Ipsi neurons
E4
ΔF/F0
Spike-rate
Spike-event
OSI
DSI

Fig. 6. 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 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 raw $\Delta F/F_0$ signal for that neuron. (**B-C**) Examples of recorded calcium signals and the predicted spike-rates and spike-events by ENS^2 (**B1-B4**, **C1-C4**). Tuning curves and selectivity indexes (**B5-B10**, **C5-C10**) are computed based on three types of inputs (**B2-B4**, **C2-C4**), respectively. (**D-E**) Comparisons of OSI/DSI computed from raw $\Delta F/F_0$ signal and after spike inference with ENS^2 for all the 192 *contra* and 141 *ipsi* neurons considered.

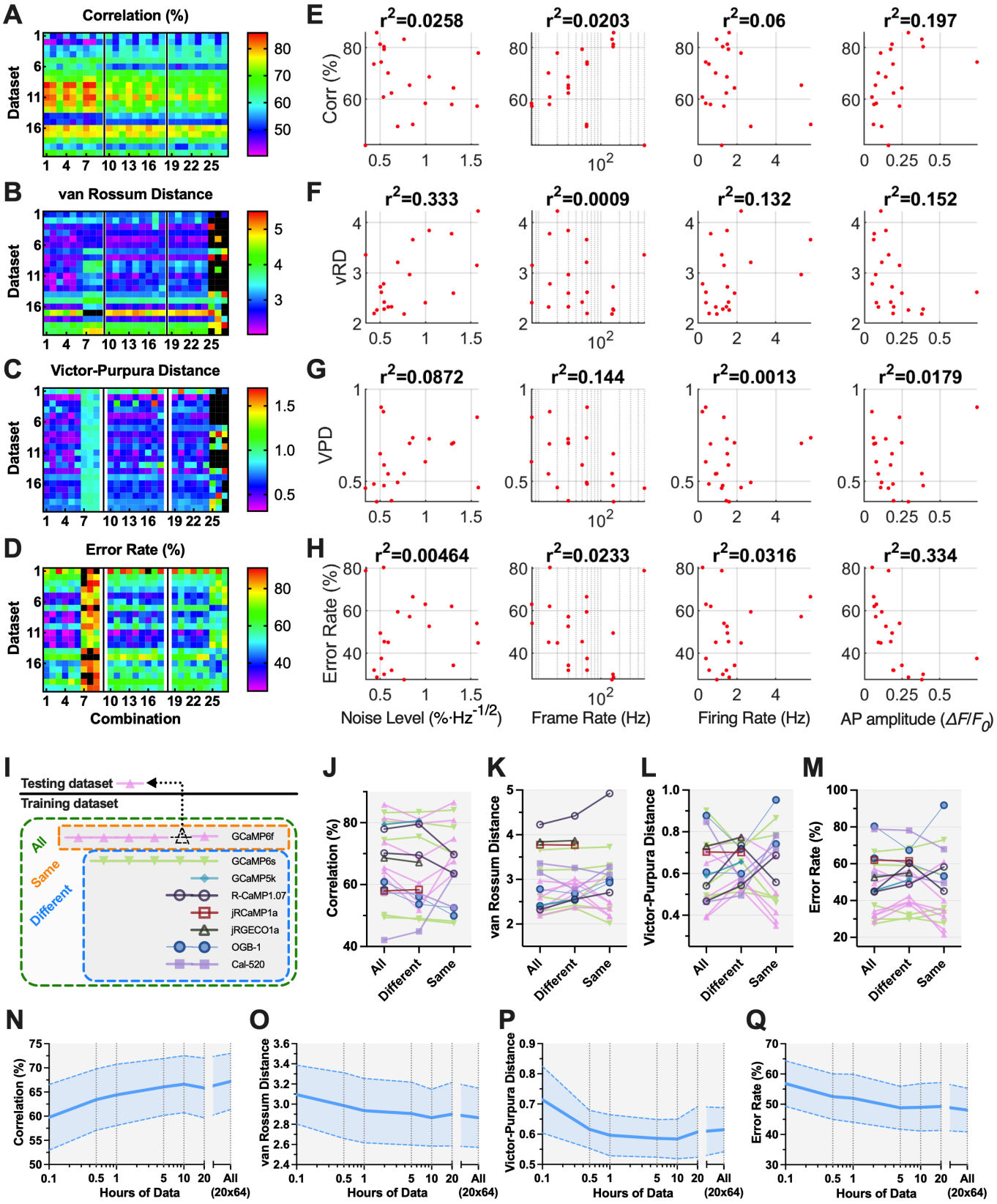

Fig. 7. Effect of dataset properties, calcium indicators, and data quantity on spike inference performance. (**A-D**) Performance of ENS[2] with different configurations (see **Fig. 2A-D**) (x-axis) on each dataset (y-axis). (**E-H**) Pearson correlation coefficients between the four properties of each dataset and the corresponding spike inference performances. (**I-M**) Performance

of spike inference for different types of calcium indicators. (**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. *All* refers to all the other 19 datasets, *Same* refers to the datasets that also used GCaMP6f and *Different* refers to all the other datasets that used calcium indicators other than GCaMP6f. (**N**-**Q**) Performance of spike inference with different length of training data. Shaded areas denote means with 95% confidence intervals.

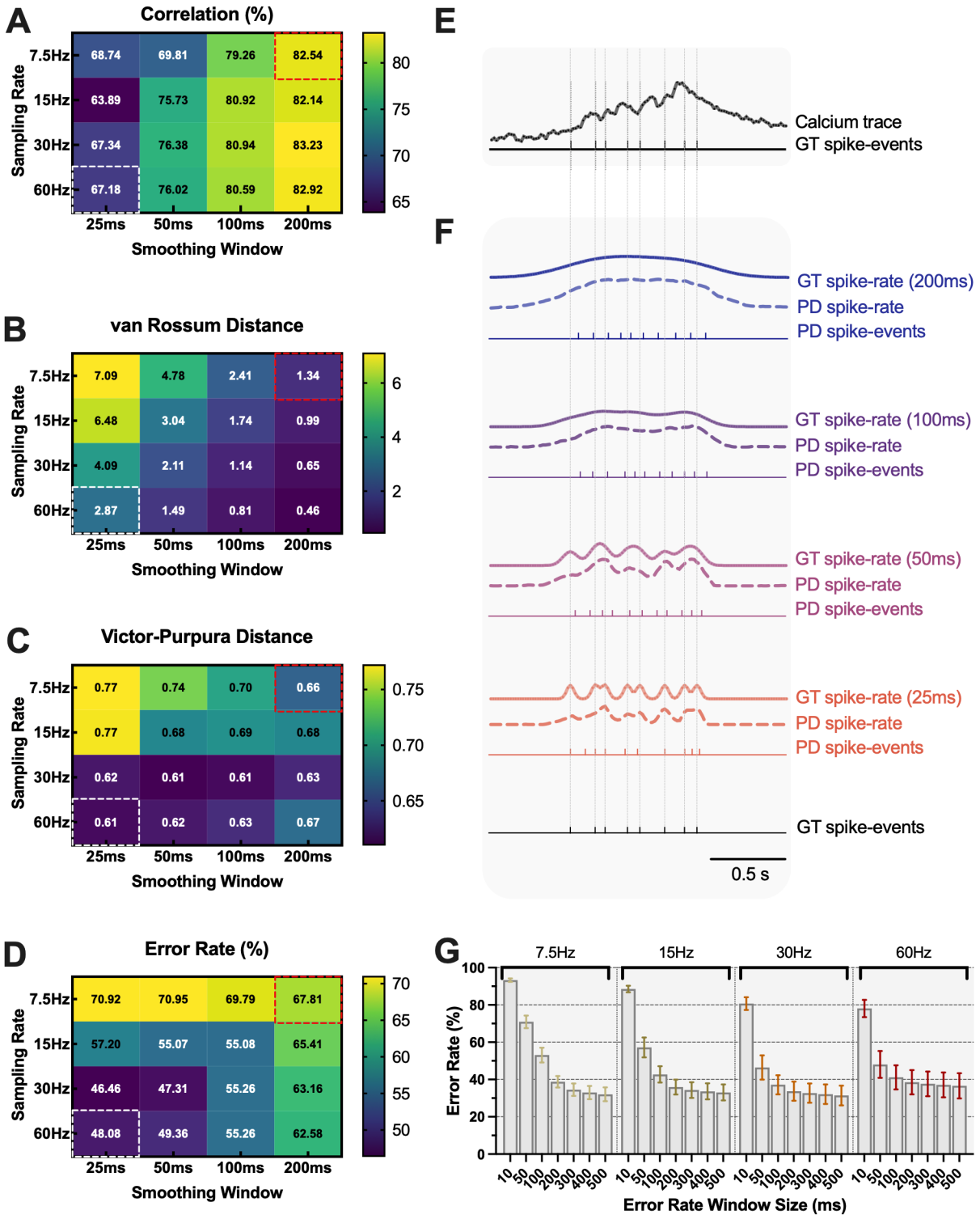

Fig. 8. Effect of evaluation metrics on spike inference performance. (**A-D**) Performance of correlation, van Rossum distance, Victor-Purpura distance, and error rate, respectively, when measured under different sampling rates and smoothing window sizes. White and red squares denote the evaluation schemes adopted by ENS[2] and CASCADE, respectively. (**E**) Example of calcium signals under 60Hz with paired spike-events. (**F**) Examples of ground truth spike-rates convolved with different smoothing window sizes (from 200ms to 25ms). The resultant spike-rate and spike-event predictions are also shown. (**G**) Performance of error rate when measured under different sampling rates and error rate window sizes.