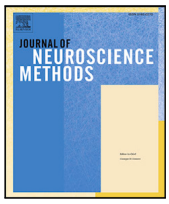
Journal of Neuroscience Methods 415 (2025) 110324

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth



STSimM: A new tool for evaluating neuron model performance and detecting spike trains similarity

A. Marasco [a,b,*], C.A. Lupascu [b], C. Tribuzi [c]

[a] Department of Mathematics and Applications, University of Naples Federico II, Naples, Italy
[b] Institute of Biophysics, National Research Council, Palermo, Italy
[c] Nova Analysis, Brescia, Italy

ARTICLE INFO

Dataset link: https://modeldb.science/2016671, https://github.com/biomemsLAB/Spike-Contrast/blob/master/Testdata.zip

Keywords:
Performance and similarity measures
Spike train analysis
Synchrony
Confusion matrix
Python

ABSTRACT

Background: In computational neuroscience, performance measures are essential for quantitatively assessing the predictive power of neuron models, while similarity measures are used to estimate the level of synchrony between two or more spike trains. Most of the measures proposed in the literature require setting an appropriate time-scale and often neglect silent periods.
New method: Four time-scale adaptive performance and similarity measures are proposed and implemented in the *STSimM* (Spike Trains Similarity Measures) Python tool. These measures are designed to accurately capture both the precise timing of individual spikes and shared periods of inactivity among spike trains.
Results: The proposed ST-measures demonstrate enhanced sensitivity over *Spike-contrast* and *SPIKE-distance* in detecting spike train similarity, aligning closely with *SPIKE-synchronization*. Correlations among all similarity measures were observed in Poisson datasets, whereas in vivo-like synaptic stimulations showed correlations only between ST-measures and SPIKE-synchronization.
Comparison of existing method: The *STSimM* measures are compared with *SPIKE-distance*, SPIKE-synchronization and *Spike-contrast* using four spike train datasets with varying similarity levels.
Conclusion: ST-measures appear more suitable for detecting both the precise timing of single spikes and shared periods of inactivity among spike trains compared to those considered in this work. Their flexibility originates from two primary factors: firstly, the inclusion of four key measures — ST-Accuracy, ST-Precision, ST-Recall, ST-Fscore — capable of discerning similarity levels across neuronal activity, whether interleaved with silent periods or solely focusing on spike timing accuracy; secondly, the integration of three model parameters that govern both precise spike detection and the weighting of silent periods.

1. Introduction

Computational models of brain regions provide a fundamental tool for investigating the dynamics of neuronal networks and the emergence of cognitive functions. However, due to technical limitations of current supercomputer systems, research has focused on efficient, simplified spiking neuron models with quantitative predictive power (see Marasco et al., 2023, 2024b,a and references therein). In these models, neuronal responses to a stimulus are represented by *spike trains*, discrete sequences of times indicating the occurrence of neuronal action potentials (spikes). Within this framework, it is crucial to assess the quality of a model's predictions by introducing performance measures to compare the model's spike trains with those generated by the original cell. Furthermore, *similarity measures* have also been used both to estimate the degree of synchrony between two or more spike trains and to quantify the reliability of neuronal responses upon repeated presentations of a stimulus. Several different approaches are reported in literature to quantify the similarity between two or more spike trains (see Victor and Purpura, 1996; van Rossum, 2001; Kreuz et al., 2011, 2013, 2015; Satuvuori et al., 2017; Cutts and Eglen, 2014). However, in most cases, similarity measures overlook the significance of detecting both the precise timing of single spikes and shared periods of inactivity among spike trains (Lyttle and Fellous, 2011). The likely reason for this limitation is that a single (parameter-free) similarity measure may not be capable of accounting for all the features of any spike train. To tackle this issue, we propose three different time-scale adaptive performance and similarity measures based on the results of the confusion matrix: ST-Accuracy, ST-Precision, ST-Recall, and additionally ST-Fscore. While these measures are conventional, their novelty lies

* Corresponding author at: Department of Mathematics and Applications, University of Naples Federico II, Naples, Italy.
E-mail address: marasco@unina.it (A. Marasco).

https://doi.org/10.1016/j.jneumeth.2024.110324
Received 20 May 2024; Received in revised form 13 November 2024; Accepted 19 November 2024
Available online 5 December 2024

Fig. 1. Logo of the *STSimM* tool.

in the mathematical approach used to determine the elements of the confusion matrix, i.e., TP (true positive), FP (false positive), TN (true negative), and FN (false negative). The ST-Accuracy measure evaluates the performance of a model or the similarity between two or more spike trains, considering both spiking and silent periods. In contrast, all other ST-measures do not account for silent periods except for mismatched spike events within these periods (FP). The proposed measures are designed to be bounded and time-scale adaptive; however, three free parameters allow for controlling both sensitivity to precise spike timing and the occurrence of silent periods. These measures are implemented in the *STSimM* tool (logo shown in Fig. 1), written in Python, and are compared to established synchrony measures *Spike-contrast* (Ciba et al., 2018), *SPIKE-distance*, *SPIKE-synchronization*, and their adaptive counterparts (Kreuz et al., 2011, 2013, 2015; Satuvuori et al., 2017).

The paper is structured as follows. A brief description of *SPIKE-distance*, *SPIKE-synchronization*, and *Spike-contrast* measures is reported in Section 2.1. Section 2.2 presents the mathematical description of the performance measures of a model and the similarity measures of two spike trains, along with simulations exploring the features of these measures. Implementation details and an overview of the four spike train datasets used for numerical comparison with these similarity measures are provided in Sections 2.3–2.4. In Section 3, we numerically compare our performance and similarity measures with the synchrony measures *SPIKE-distance*, *SPIKE-synchronization*, and *Spike-contrast* on four train datasets. Finally, Section 4 contains the discussion and concluding remarks.

2. Material and methods

2.1. SPIKE-distance, SPIKE-synchronization, and Spike-contrast measures

In this section, we provide a brief overview of the similarity measures *SPIKE-distance*, *SPIKE-synchronization*, and *Spike-contrast*. For a detailed derivation of these measures, as well as their mathematical properties, the reader is referred to Kreuz et al. (2011, 2013, 2015), Satuvuori et al. (2017), Ciba et al. (2018) and Mulansky et al. (2015).

To quantify the degree of (dis)similarity and synchronization of two (or more) spike trains two parameter-free measures have been introduced in Kreuz et al. (2011, 2013, 2015), namely *SPIKE-distance* and *SPIKE-synchronization*. In particular, *SPIKE-distance* is a measure of dissimilarity that yield the value zero for identical spike trains, while *SPIKE-synchronization* is a measure of similarity with high values denoting similar spike trains. In *SPIKE-distance*, the dissimilarity profile $S(t)$ in a fixed interval $[0, T]$ is calculated in two steps: first for each spike a spike time difference is calculated and then for each time instant the relevant spike time differences are selected, weighted, and normalized. Then, the overall distance value D_S is defined as temporal average of the dissimilarity profile as follows

$$D_S = \frac{1}{T}\int_0^T S(t)\,dt. \tag{1}$$

SPIKE-synchronization is complementary to *SPIKE-distance*, since it is a measure of spike matching based on a binary coincidence criterion. The coincidence detection uses a coincidence window, which denotes the time lag below which two spikes from two different spike trains are considered to be coincident. The coincidence windows are adapted to the local spike rates, and the coincidence criterion is quantified by means of a coincidence indicator for each individual spike of the two spike trains, assigning either a one or a zero to each spike depending on whether it is part of a coincidence or not. Then, to obtain one combined similarity profile, the authors pool the spikes of the two spike trains as well as their coincidence indicators by introducing one overall spike index k. This yields one unified set of coincidence indicators $C(t_k)$, in which each coincidence leads to a pair of consecutive ones. Finally, *SPIKE-synchronization* is defined as average value of this profile

$$S_C = \frac{1}{M}\sum_{k=1}^{M} C(t_k), \tag{2}$$

where M denotes the total number of spikes in the pooled spike train.

In summary, *SPIKE-synchronization* quantifies similarity instead of difference as *SPIKE-distance* does. However, *SPIKE-distance* can be converted from a measure of distance into a measure of similarity by considering $1-D_S$. In Satuvuori et al. (2017), these measures have been adapted for data containing multiple timescales by adding a notion of the relative importance of local differences compared to the global timescales. These generalizations, called *A-SPIKE-distance* and *A-SPIKE-synchronization*, are built on a minimum relevant time scale (MRTS) which is implemented via the threshold parameter $\mathcal{T}$. These generalized measures fall back on the original definitions when $\mathcal{T} = 0$.

In the following, we will compare the proposed measures implemented in *STSimM* Python tool with the similarity measures $1 - D_S$ and S_C defined by Eqs. (1)–(2) and corresponding to *SPIKE-distance* and *SPIKE-synchronization*, respectively, as well as with their adaptive counterparts.

In Ciba et al. (2018) a time-scale independent spike train synchrony measure called *Spike-contrast* is proposed. The synchrony measure S over a signal length T is defined as the maximum of synchrony function $s(\Delta)$ over different bin sizes Δ

$$S = \max_{\Delta} s(\Delta), \quad \Delta \in [\Delta_{min}, \Delta_{max}], \tag{3}$$

where $\Delta_{min} = \max(\mathrm{ISI}_{\min}, L), \Delta_{max} = T/2$, with $\mathrm{ISI}_{\min}$ being the minimum value of the ISIs in the two spike trains and L a suitable constant value. If L is set to zero, the measure S is adaptive to the data.

There are alternative approaches that address the challenge of analyzing data across multiple timescales. For instance, Lyttle and Fellous (2011) introduced a measure specifically designed to evaluate the similarity of spike trains, with a particular focus on detecting bursts and common silent periods. This approach requires the adjustment of two timescale parameters and three additional factors: the minimum duration of silent periods, the duration of burst ISIs, the minimum number of spikes in a burst, a scaling factor to weigh the importance of bursts relative to single spikes, and another factor to balance the detection of bursts and silent periods. Although this extensive parametrization provides the experimenter with a powerful and flexible analytical tool and greater control over the analysis, it can pose challenges, especially when dealing with high-dimensional data. In this work, we will not compare our measures with those reported in Lyttle and Fellous (2011) solely because the numerical code associated with this approach is not available.

2.2. Mathematical framework

In our context, neuronal responses to a stimulus are represented by a spike train, a discrete sequence of times indicating the occurrence of neuronal action potentials (spikes). In the following, we introduce a set of measures that quantify the degree of similarity between two or more spike trains. These performance measures were previously used to assess the quality of predictions of the A-GLIF (Adaptive generalized leaky integrate-and-fire) model in reproducing spike trains in response to a large set of synaptic stimulations within a given time interval (Marasco et al., 2024b). However, as will be shown below, they can be extended to quantify both the similarity and synchronization of two or more spike trains through a symmetrization procedure.

2.2.1. Performance measures of predictive capability of a neuronal model

We will begin by describing the performance measures employed to accurately identify the correspondence of spike times and silent periods

between the experimental and model spike trains. Let

$$\mathrm{X}_{\mathrm{exp}} = \{t_1, \dots, t_N\}, \qquad \mathrm{X}_{\mathrm{mod}} = \{\bar{t}_1, \dots, \bar{t}_L\}, \tag{4}$$

be the spike trains generated by the cell and predicted by a (spiking neuron) model, respectively.

To verify whether a model correctly identifies the presence (or absence) of a spike within a given time interval $[t_0, T]$, we resort to the following *performance measures* between the spike trains $\mathrm{X}_{\mathrm{exp}}$ and $\mathrm{X}_{\mathrm{mod}}$

$$\begin{aligned} m_A(\mathrm{X}_{\mathrm{exp}}, \mathrm{X}_{\mathrm{mod}}) &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}, && (\mathrm{ST}-\mathrm{Accuracy}) \\ m_P(\mathrm{X}_{\mathrm{exp}}, \mathrm{X}_{\mathrm{mod}}) &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, && (\mathrm{ST}-\mathrm{Precision}) \\ m_{\mathrm{R}}(\mathrm{X}_{\mathrm{exp}}, \mathrm{X}_{\mathrm{mod}}) &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, && (\mathrm{ST}-\mathrm{Recall}) \\ m_{\mathrm{F}}(\mathrm{X}_{\mathrm{exp}}, \mathrm{X}_{\mathrm{mod}}) &= \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, && (\mathrm{ST}-\mathrm{Fscore}) \end{aligned} \tag{5}$$

where TP (True Positives) is the number of spikes from the experimental trace that are also found in the model; TN (True Negatives) is the number of intervals in which the neuron does not fire a spike in both the experimental trace and the model; FP (False Positives) is the number of mismatched spikes in the model; FN (False Negatives) is the number of spikes in the experimental trace that are not matched in the model. All the quantities in (5) represent probability measures within a given time interval; therefore they are bounded and their values may depend on the number of events occurring during simulations of different lengths. The *ST-Accuracy* defines how accurately the model reproduces both spiking and silent periods in an experimental trace. A value of 1 indicates an ideal scenario of a perfect correspondence of spike times and pause between the cell and the model. Furthermore, all the other performance measures (5) do not consider silent periods, except for mismatched spike events within these periods (FP). The *ST-Precision* represents the percentage of relevant spike times (TP) among all the retrieved spike times (TP+FP), whereas *ST-Recall* is the fraction of correct detections (TP) over all those generated by the cell (TP+FN). Finally, the *ST-Fscore* is the harmonic mean of the ST-Precision and ST-Recall measures. Similarly to other measures, the F-score also takes value between 0 and 1, where 1 indicates the optimal scenario in which neither false positives nor false negatives are detected.

Let N be the number of spike in the recorded experimental trace in the time window $[t_0, T]$. For any spike in the experimental trace at time t_i, TP, FP, and FN are calculated by exploring the behavior of the model in the following interval (see gray regions in Figs. 2–4)

$$I(t_i) = \left[t_i - \omega(t_i - t_{i-1}), t_i + \omega(t_{i+1} - t_i)\right], \quad i = 1, \dots, N \tag{6}$$

where $t_{N+1} = T$ and $0 < \omega \le 0.5$ to avoid overlaps between consecutive intervals (6), in such a way that

$$\mathrm{TP} = \begin{cases} 1 & \text{if there is at least 1 spike in the model,} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

$$\mathrm{FP} = \begin{cases} n-1 & \text{if there are } n > 1 \text{ spikes in the model,} \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

$$\mathrm{FN} = \begin{cases} 1 & \text{if there are no spikes in the model,} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

We remark that, in contrast with the conventional way to calculate these quantities, here we are using a variable interval (6), which is adapted to follow the instantaneous firing behavior of the cell that can be characterized by periods of intense or moderate activity, eventually interleaved with silent periods. Consequently, intervals between consecutive spikes, i.e., inter-spike intervals (ISIs), can vary greatly in duration, being very short, moderate, or long. To accommodate long ISIs, the maximum amplitude $A_i = \omega(t_{i+1} - t_{i-1})$ of any interval (6) was set to 2Λ. In detail, when $\omega(t_i - t_{i-1})$ and/or $\omega(t_{i+1} - t_i)$ are greater than Λ, we modify the intervals (6) as follows (see Fig. 2)

$$I(t_i) = \left[t_i - \varphi_i, t_i + \varphi_{i+1}\right], \quad i = 1, \dots, N \tag{10}$$

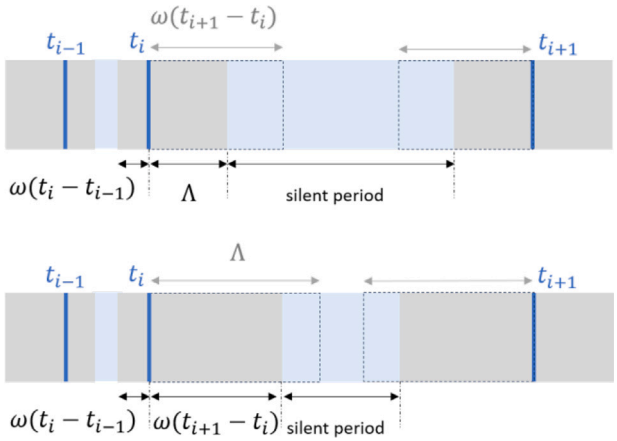


Fig. 2. Determination of the intervals $I(t_i)$ and $\tilde{I}(t_i)$. (top) If $\omega(t_i - t_{i-1}) > \Lambda$, the interval $I(t_i)$ is set according to (10) where $\varphi_i = \Lambda$. (bottom) For $\omega(t_i - t_{i-1}) < \Lambda$, the interval $I(t_i)$ follows (6). Blue vertical bars represent spike times; gray and light-blue regions represent the intervals $I(t_i)$ and $\tilde{I}(t_i)$, respectively.

where

$$\varphi_i = \begin{cases} \omega(t_i - t_{i-1}), & \text{if } \omega(t_i - t_{i-1}) \le \Lambda, \\ \Lambda, & \text{otherwise,} \end{cases} \tag{11}$$

and, according to the neuroscience literature, Λ could range from 10 ms to 1 s (see Cutts and Eglen, 2014).

To calculate TN and any other FP, we consider the following intervals where there is no activity in the cell (see light blue regions in Figs. 2–4)

$$\begin{aligned} \tilde{I}(t_0) &= \left[t_0, t_1 - \omega(t_1 - t_0)\right], \\ \tilde{I}(t_i) &= \left[t_i + \omega(t_{i+1} - t_i), t_{i+1} - \omega(t_{i+1} - t_i)\right], \quad i = 1, \dots, N-1, \\ \tilde{I}(t_N) &= \left[t_N + \omega(t_{N+1} - t_N), t_{N+1}\right]. \end{aligned} \tag{12}$$

Owing to Eq. (11), the intervals $\tilde{I}(t_i)$ become (see light blue regions in Fig. 2)

$$\begin{aligned} \tilde{I}(t_0) &= \left[t_0, t_1 - \varphi_1\right], \\ \tilde{I}(t_i) &= \left[t_i + \varphi_{i+1}, t_{i+1} - \varphi_{i+1}\right], \quad i = 1, \dots, N-1, \\ \tilde{I}(t_N) &= \left[t_N + \varphi_{N+1}, t_{N+1}\right]. \end{aligned} \tag{13}$$

Finally, to evaluate the FP relative to the intervals (13) we use the following rule

$$\mathrm{FP} = \begin{cases} n & \text{if there are } n \text{ spikes in the model,} \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

whereas TN is equal to the number S_i of subintervals of $\tilde{I}(t_i)$, of maximum amplitude Λ_i, without any spike in the model. In order not to overestimate the number of TNs in long silent intervals, the number S_i and the parameter Λ_i are determined as follows

$$S_i = \begin{cases} 1, & \text{if } \tilde{A}_i \le 2\Lambda, \\ \left\lceil \dfrac{\tilde{A}_i}{2\Lambda} \right\rceil, & \text{if } 2\Lambda < \tilde{A}_i < 2c\,\Lambda, \\ \lceil c \rceil, & \text{otherwise,} \end{cases} \qquad \Lambda_i = \frac{\tilde{A}_i}{S_i}, \tag{15}$$

where $\tilde{A}_i$ is the amplitude of each interval (13), $\lceil \cdot \rceil$ denote the ceiling function, and $c \ge 1$ is a suitable positive constant such that it results $1 \le S_i \le \lceil c \rceil$ for all $i = 0, \dots, N$ (see light blue regions in Figs. 2 and 3).

As depicted in Fig. 3, the performance measures (5) are not symmetric, i.e. $m_i(\mathrm{X}_{\mathrm{exp}}, \mathrm{X}_{\mathrm{mod}}) \neq m_i(\mathrm{X}_{\mathrm{mod}}, \mathrm{X}_{\mathrm{exp}})$ for $i = A, P, R, F$. In this context, the objective is to evaluate how accurately the model replicates experimental data. Consequently, the experimental spike train X_{exp} (red bars in Fig. 3) serves as the reference, and our focus lies in assessing the extent to which the modeled spike train X_{mod} (blue bars

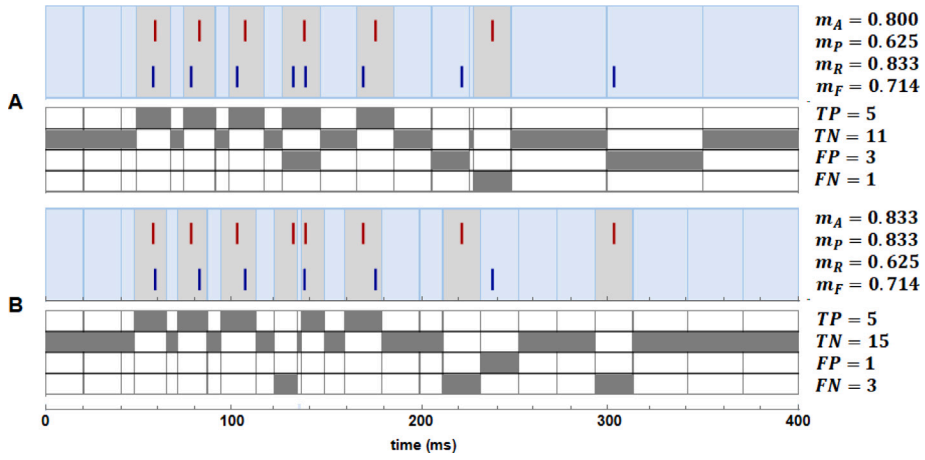


Fig. 3. Illustration of the calculation procedures of the performance metrics (5). (Panels A and B, top) Raster plots representing spike times from experiments (red bars) and the model (blue bars). Light blue and gray regions refer to intervals (6) and (12), respectively, for $\Lambda = 10$ ms, $\omega = 0.35$, and $c = 3$. (Panels A and B, bottom) The spike trains were scanned to test for mismatch of spikes or silent periods occurring within a variable time window.

in Fig. 3) faithfully reproduces both spike timings and silent intervals. For instance, in panel A of Fig. 3 the experimental spike train exhibits 6 spikes, while the spike train generated by the model contains 8 spikes. With parameter choices $\Lambda = 10$ ms, $\omega = 0.35$, and $c = 3$, it is observed that the train X_{mod} accurately reproduces only 5 spikes (those within the gray intervals), while the remaining 3 (one in the fourth gray interval and the other two within the light blue intervals) are false positives. Regarding the silent periods, both spike trains share all but those in which the model generates 2 false positive spikes. Conversely, in panel B of Fig. 3, the roles of the two spike trains are reversed, with the experimental one now containing 8 spike times while the model generated one has only 6. With the same parameter choices, the train X_{mod} accurately reproduces 5 spikes (those within the gray intervals), while the sixth one is a false positive. However, in this case the two spike trains share all silent intervals except for the one where the false positive occurs. The two experimental spike trains depicted in Fig. 3 exhibit dynamics different from what is reproduced by the model. This disparity is underscored by the varying values of performance metrics in the two cases. For instance, in the second case, the ST-Accuracy is higher because the model shares a greater number of silent intervals with the experimental data (11 TN in the first case and 15 in the second). Conversely, ST-Precision and ST-Recall assume exchanged values in the second case compared to the first, resulting in an unchanged value of ST-Fscore.

We note that the ST-Accuracy m_{A} depends on all parameters ω, Λ, and c, whereas the other performance measures in (5) depend solely on ω and Λ (see Fig. 4). In detail, as shown in Fig. 4, we examined the effects of varying each of the parameters ω, Λ, and c on performance measures (5). The parameter ω affects the amplitude of the intervals $I(t_i)$ of Eq. (10) where, for each experimental spike time t_i, we search for the corresponding spike time in the train generated by the model. In particular, the amplitude of each interval $I(t_i)$ monotonically increases with respect to ω (light gray intervals in panels A and B of Fig. 4). Similarly, the parameter Λ controls the maximum amplitude of all intervals $I(t_i)$ (light gray intervals in panels A and C of Fig. 4). Furthermore, the parameter c represents the maximum number of sub-intervals in each silent period $\tilde{I}(t_i)$ in Eq. (13) (light blue intervals in panels A and D of Fig. 4).

In neuroscience, timescales typically range from milliseconds to seconds, with shorter timescales generally deemed irrelevant (Cutts and Eglen, 2014; Marasco et al., 2024b). Appropriate timescales clearly depend on the specific system being studied. In fields such as economics, engineering, geology, and other research areas, the relevant timescales can range from hours and days to even months and years. In these cases, setting the appropriate values of the parameters ω, Λ, and c for a given dataset might not be a simple task, especially when different timescales are relevant. To address this issue, we propose an automatic selection of the values for these parameters (see panel E of Fig. 4). In particular, we set ω and c to their maximum and minimum values, respectively, while for Λ we adopt a formula analogous to that reported in Satuvuori et al. (2017) to account for the different timescales present in the data, i.e.

$$\omega = 0.5, \quad \Lambda = \frac{1}{4}\sqrt{\frac{\sum_{h=1}^{N} |\text{ISI}_h|^2}{N}}, \quad c = 1, \tag{16}$$

where N is the total number of ISIs in the experimental spike train X_{exp}.

The user of the *STSimM* tool can choose the automatic setting for each parameter. Moreover, unlike ω and c, the parameter Λ represents a time and has the same unit of measurement as the time series under consideration. To provide greater control over the results, when launched with automatic parameters, the *STSImM* tool will also output the value of Λ determined by using Eq. (16). This allows the user to rerun the Python tool with a more appropriate value for Λ if the method identifies a value that is not significant for the time series under analysis.

For the spike trains shown in Fig. 4, the value of Λ determined by Eq. (16) is 9.6 ms, which is entirely acceptable for the data under analysis. However, it is easy to show cases where the value of Λ automatically determined by the *STSimM* tool falls completely out of range, making it necessary to rerun the Python tool with a manually selected, more appropriate value of Λ.

For instance, let

$$X_{\text{exp}} = \{100, 3900\}, \qquad X_{\text{mod1}} = \{2200, 4500\}, \qquad X_{\text{mod2}} = \{1500, 4500\} \tag{17}$$

be the spike trains generated by a cell (X_{exp}) and predicted by two different point-neuron models (X_{mod1}, and X_{mod2}), respectively, in the time interval $[0, 8000]$ ms. When we analyze the spike train pairs $(X_{\text{exp}}, X_{\text{mod1}})$ and $(X_{\text{exp}}, X_{\text{mod2}})$ shown in panels A and B of Fig. 5, respectively, the results obtained using the *STSimM* tool with $\omega = 0.35$, $\Lambda = 10$ ms, and $c = 3$ differ significantly from those obtained with the automatic selection of the same parameters. In these cases, the automatic values of Λ, determined by using Eq. (16), are 785.2 ms and 855.8 ms, respectively. Although for $(X_{\text{exp}}, X_{\text{mod1}})$, the *ST-Fscore* value for the automatic selection of the parameters is identical to that obtained with *SPIKE-synchronization* and *A-SPIKE-synchronization*, the two spike trains cannot be considered either similar or synchronized. Furthermore, despite a slight improvement in the distance between the first spike times in the two trains $(X_{\text{exp}}, X_{\text{mod2}})$ (see Fig. 5B), unlike

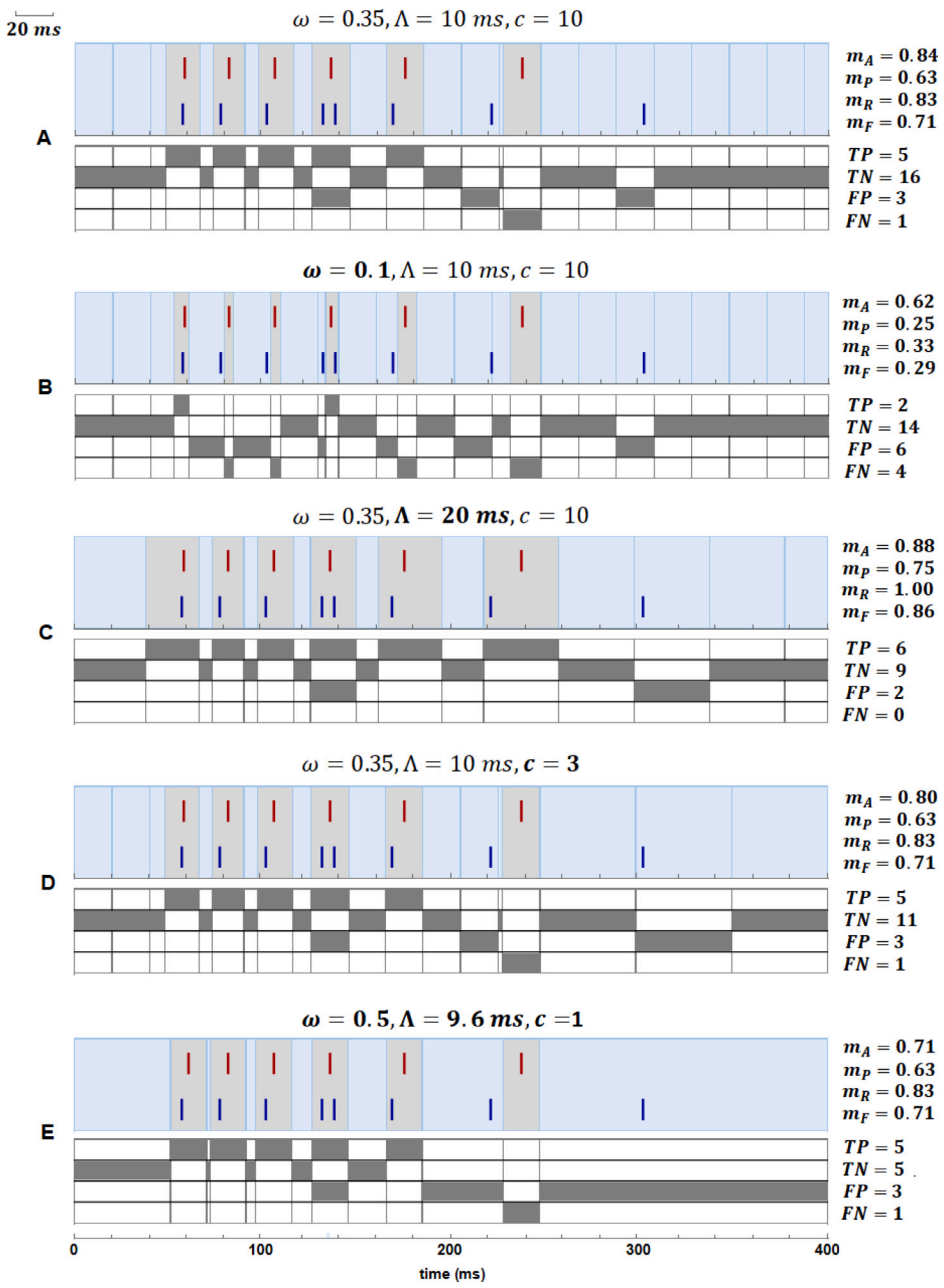


Fig. 4. Parameter influence on performance measure estimation. (top) Raster plots representing spike times from experiments (red bars) and the model (blue bars). (A, B) Effects of changes in parameter ω affecting the amplitude of the intervals $I(t_i)$; (A, C) Effects of changes in parameter Λ controlling the maximum amplitude of the intervals $I(t_i)$; (A, D) Effects of changes in parameter c determining the number of sub-intervals of $\tilde{I}(t_i)$. (A, E) Effect of automatic selection of parameters ω, Λ, and c. All performance measures (5) are monotonically increasing w.r.t. to each parameter ω, Λ and c.

the *ST* and *Spike-contrast* measures, all the other metrics show higher values compared to the case reported in panel A, and even for *SPIKE-synchronization* and *A-SPIKE-synchronization*, the maximum value is reached.

These simple examples show how the automatic selection of the parameter Λ, which represents the maximum half-width of each interval in which two spike times are considered synchronous, can lead to completely incorrect results. However, knowing the automatically calculated value of Λ can easily allow users to make a more appropriate choice for the phenomenon under examination. In fact, by using $\Lambda = 10$ ms, the value of the *ST-Fscore* is always zero, clearly showing that neither model is able to accurately reproduce the experimental spike timings (see the middle and bottom figures in Panels A and B of Fig. 5).

2.2.2. Similarity measures of two or more spike trains

As previously stated, the performance measures (5) are not symmetric. Furthermore, these measures are not suitable for quantifying the reliability of neuronal responses upon repeated presentations of a stimulus in a cell or for estimating the degree of similarity between two or more spike trains across different cells. In fact, in this case, we need to determine a set of measures that are invariant with respect to the order in which each pair of spike trains is analyzed. Then, to quantify the overall similarity of two (or more) spike trains, it is necessary to modify the measures in (5) to make them symmetric. To this end, we introduce a set of *similarity measures* by imposing that each of these quantities is equal to the mean of the directed measures as reported in (5).

In detail, let X_i, X_j be two spike trains, then we define

$$M_H(X_i, X_j) = \frac{1}{2}(m_H(X_i, X_j) + m_H(X_j, X_i)), \quad H = A, P, R, F \tag{18}$$

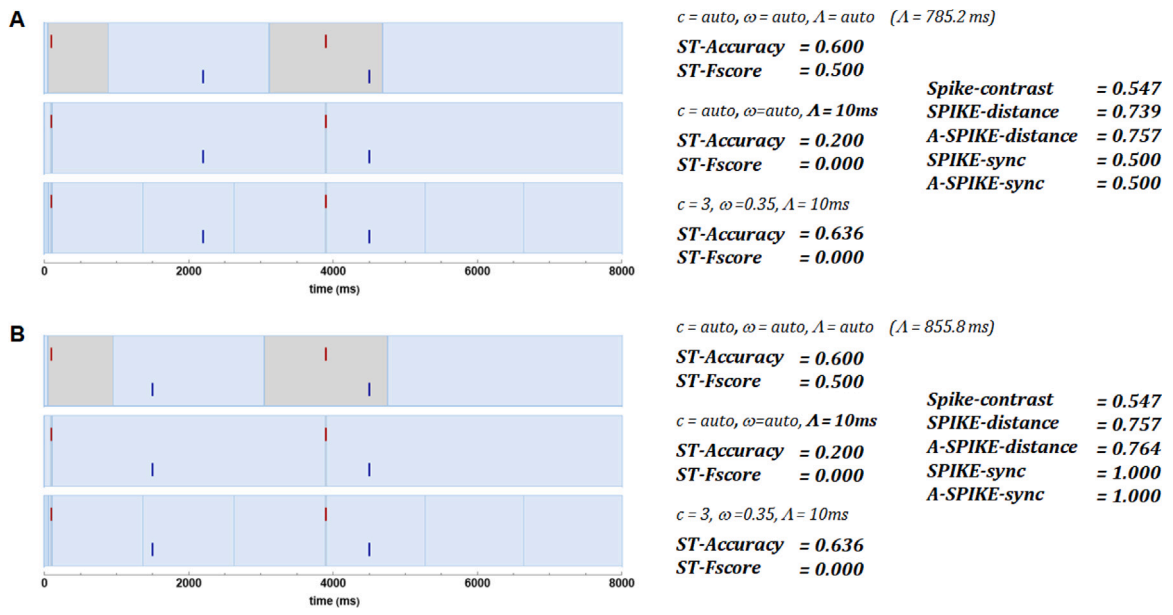


Fig. 5. Influence of automatic parameters selection on the similarity of spike trains. Comparison of spike trains $(X_{\text{exp}}, X_{\text{mod1}})$ and $(X_{\text{exp}}, X_{\text{mod2}})$ using automatic and manual selection of the parameters ω, Λ, and $c = 3$. Panels A and B show the results for different parameter configurations, along with their respective similarity and synchronization measures for the spike train pairs $(X_{\text{exp}}, X_{\text{mod1}})$ and $(X_{\text{exp}}, X_{\text{mod2}})$, respectively.

where m_H is any of the measures described in (5), and M_H the corresponding symmetric counterpart.

Furthermore, to quantify the overall similarity of n spike trains $\mathrm{X}_1, \ldots, \mathrm{X}_n$ we resort to the following straightforward extension

$$M_H(\mathrm{X}_1, \ldots, \mathrm{X}_n) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} M_H(\mathrm{X}_i, \mathrm{X}_j), \quad H = A, P, R, F. \tag{19}$$

In panels A and B of Fig. 3, the performance measures of the spike trains $(X_{\text{exp}}, X_{\text{mod}})$ and $(X_{\text{mod}}, X_{\text{exp}})$ are presented, respectively. In this scenario, the similarity measures (18) yield $M_A = 0.815$, $M_P = M_R = 0.729$, and $M_F = 0.714$.

Also in this case, the automatic selection of the parameters ω, Λ, and c follows Eq. (16), provided that N is defined as the sum of the total number of ISIs of the two spike trains X_i and X_j, for all pairs (i, j).

2.3. Measures implementation

The *STSimM* measures are implemented in Python and can be cloned and executed as a Python script directly from ModelDB section of the Senselab database (https://modeldb.science/2016671). The *STSimM* tool is open-source under the BSD 3-Clause license, allowing the use, redistribution, and modification of both its source and binary formats. The performance and similarity measures described in Sections 2.2.1–2.2.2 were computed in parallel using the multiprocessing module in Python. Moreover, to speed up the Python code, we optimized it using the Cython library. Since all Python code is completely valid in Cython, we simply cythonized the Python code by generating the associated .pyx file. This resulted in a significant reduction in computational time, as Cython compiles the code to machine-level instructions specific to the operating system, thus eliminating the requirement for a CPython interpreter.

The *STSimM* tool allows to compute both the performance measures (5) by choosing the input parameter `weight=1` and the similarity measures (18) by setting `weight=0.5` (see Fig. 6).

As shown in Fig. 6, the input parameters of the *STSimM* tool include: two lists of data, `train1` and `train2`, representing the spike trains to be analyzed; the `weight` variable, which specifies the weight applied to the first train (with the weight of the second train calculated as `1-weight`); the values of the model parameters ω, Λ, and c; and the maximum stimulation time, `max stimulation time`. The start time of the simulation is set to 0 ms. The values of the model parameters ω, Λ, and c can be specified by the user or set to `auto` to use default values as described in Eq. (16). When `lambda=auto` is used, the Python script outputs the resulting Λ value for the analyzed spike trains. When using the *STSimM* tool to compute performance measures against experimental data, note that `train1` must correspond to the experimental spike train, and the `weight` parameter must be set to 1. The tool outputs ST-Accuracy, ST-Precision, ST-Recall, and ST-Fscore. If a measure cannot be calculated due to the nature of the spike train pairs, its value is returned as `None`.

2.4. Test datasets

To evaluate the *STSimM* tool with respect to the *SPIKE-distance*, *SPIKE-synchronization*, and *Spike-contrast* measures, we utilize four spike train datasets with different similarity or synchrony levels. Specifically, we employ spike train dataset obtained from in vivo-like synaptic stimulations (see Marasco et al., 2024b) to compare the performance measures (5) on the A-GLIF models with those obtained using *SPIKE-distance*, *SPIKE-synchronization*, and *Spike-contrast*. This spike train dataset is available for download at the ModelDB section of the Senselab database (https://modeldb.science/2016671). Additionally, we assess the similarity measures (18) on three synthetic datasets generated from Poisson spiking, bursting, and sub-bursting models as reported in Cutts and Eglen (2014) and Ciba et al. (2018). The three synthetic datasets are available for download at https://github.com/biomemsLAB/Spike-Contrast/blob/master/Testdata.zip.

2.4.1. Synaptic in-vivo like stimulations

In Marasco et al. (2024b), a set of 130 synaptic stimulation currents for a CA1 pyramidal neuron was generated using a double exponential current (with rise and decay time constants of 0.5 ms and 20 ms, respectively), activated by means of the spike times recorded during open-field explorations from 10 CA3 pyramidal neurons of male Long-Evans rats. Furthermore, to model the physiologically plausible variability in the synaptic strength, peak synaptic currents in the range of 50–1250 pA with a step of 100 pA were used. A representative spike train pair of the somatic voltage in the reference CA1 pyramidal neuron and A-GLIF models is reported in panel A of Fig. 8.

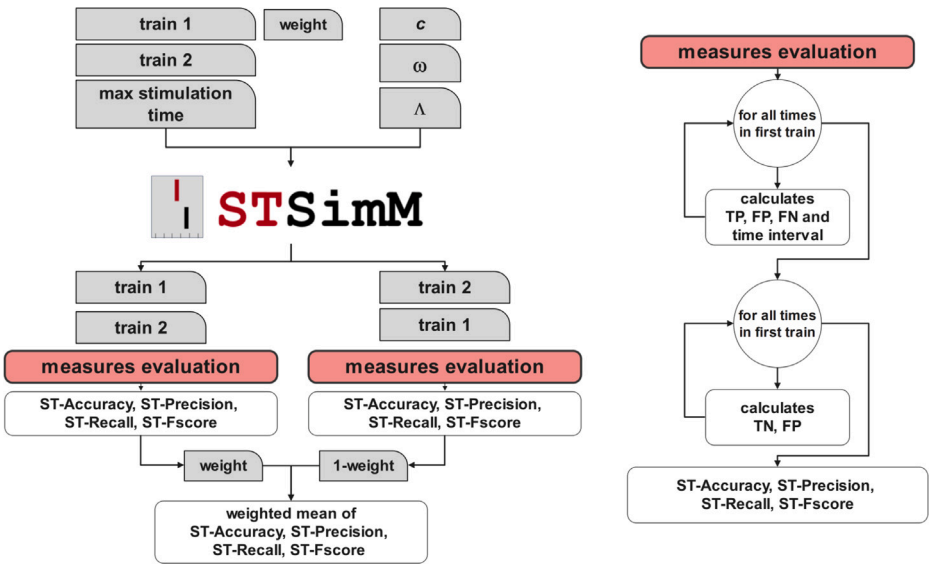


Fig. 6. Flow-chart illustration of the *STSimM* algorithm. The input data includes (i) two spike trains, (ii) the value of weight 1, and (iii) the values of the parameters ω, Λ, and c. The tool provides the values of the performance measures (5) when `weight=1` and the similarity measures (18) when `weight=0.5`.

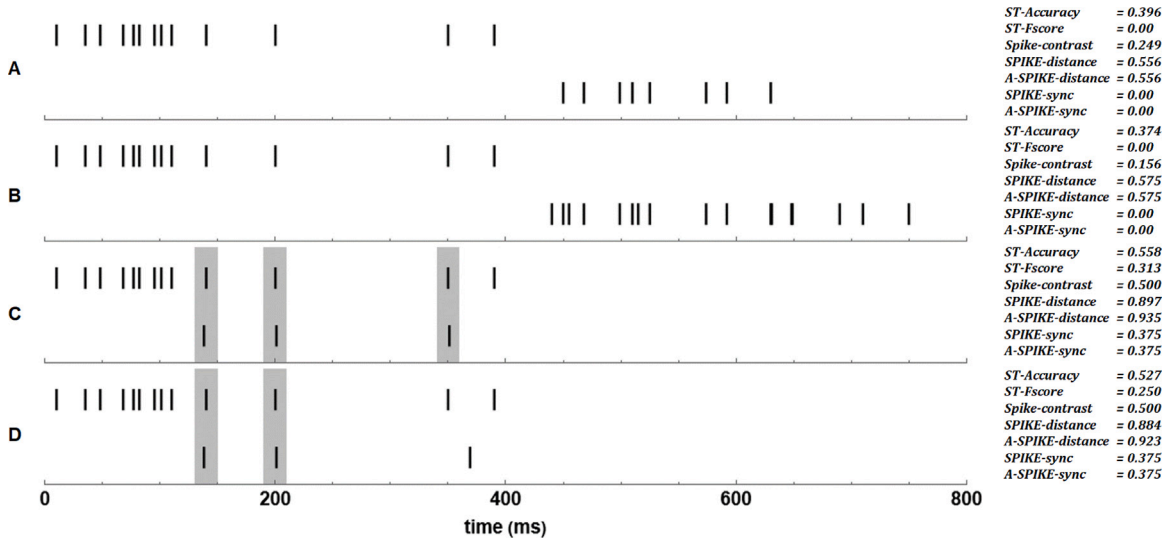


Fig. 7. Comparison of the similarity measures. (A) Two spike trains completely desynchronized with 13 and 8 spike events, respectively; (B) As in panel A but with the second train having 15 spikes; (C) At the top the same spike train as in panels A and B, while the second presents only 3 spike events. (D) Same as panel C but with the last spike event of the second train shifted by about 20 ms. In all panels, the similarity measure (18) are computed by setting $\omega = 0.35$, $\Lambda = 10$ ms, and $c = 3$. On the right of all panels the similarity measure (18) are compared with those obtained by *Spike-contrast*, *SPIKE-distance*, *SPIKE-synchronization*, and their adaptive counterparts.

2.4.2. Poisson spike model

First, we consider two spike trains X_1 and X_2 generated by a Poisson process with a spike rate λ of 1.5 spikes per second and a signal length of 300 s as described in Cutts and Eglen (2014). Both spike trains share a defined fraction of synchronous spikes from a Poisson process with a rate of

$$\lambda_s = \lambda\,(1 - F), \tag{20}$$

where F is a factor that ranges from 0 to 1 with a step of 0.05, defining different levels of synchrony (see Ciba et al., 2018). Finally, for each F, $n = 20$ spike train pairs were generated. In the left plots of Fig. 9A, two spike train pairs with high ($F = 0$) and low ($F = 1$) synchrony are shown.

2.4.3. Poisson burst model

A set of data that replicates bursts of spikes is generated by a doubly stochastic process. Firstly, the position of the burst is determined as for the spike times in Section 2.4.2 with a shared fraction λ_s of burst center, and $\lambda = 0.05$ burst per second as in Eq. (20). Then, the number of spikes in the burst and a position for each spike are generated (see Ciba et al., 2018 for further details). Also in this case, for each F, $n = 20$ spike train pairs were generated. As illustrated in Fig. 9B (left) even when we consider a dataset with high synchrony ($F = 0$) the spike trains in a single burst are not identical.

2.4.4. Poisson sub-bursts model

This dataset refers to bursts of spikes containing shorter bursts, i.e. sub-burst. Following the procedure described in Ciba et al. (2018), a spike train X_1 containing sub-bursts made of spikes with ISIs of 0.02 s was generated. Then, a second spike train X_2 was created from X_1 with each spike subjected to a random jitter, i.e. small perturbation of the spike times. The amount of random jitter was drawn from a uniform distribution in the range $[0\text{ s}, 0.02\text{ s}]\ F$. Following this procedure, when $F = 0$ no jitter was applied, resulting in identical spike trains. Then,

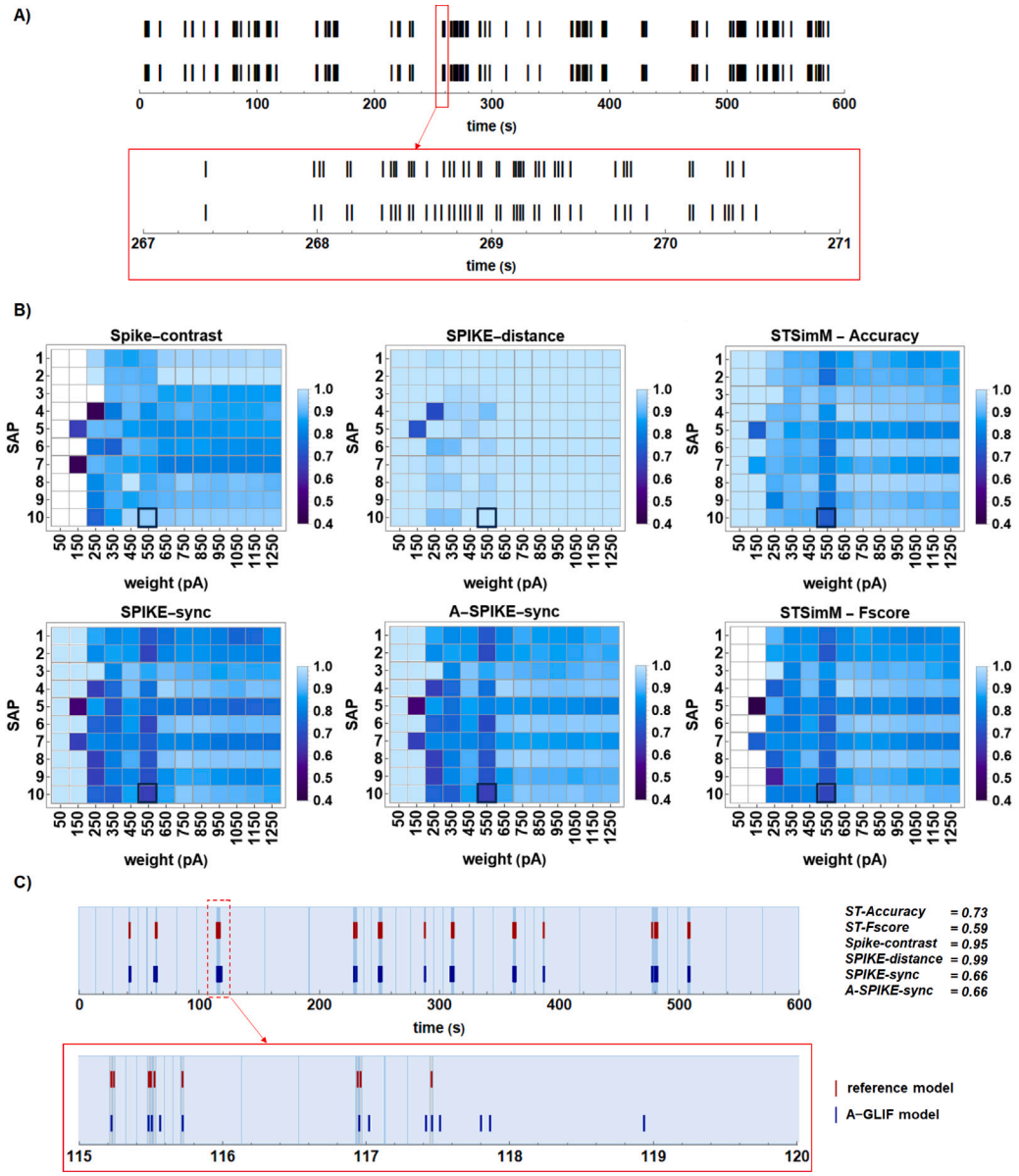


Fig. 8. Synaptic in-vivo like stimulations. A. (top) Reference (train 1) and A-GLIF model (train 2) raster plots generated by the synaptic stimulation SAP 7 with a peak synaptic current of 650 pA during a 10 min stimulation; (bottom) Same as for the top but in the time interval 267–271 s. B. Comparison of the performance measures *ST-Accuracy* and *ST-Fscore* (right) with *Spike-contrast* (left, top), *SPIKE-distance* (middle, top), *SPIKE-synchronization* (left, bottom), *A-SPIKE-synchronization* (middle, bottom) measures, for the references and A-GLIF model traces obtained in response to 130 synaptic stimulations. White regions refer to traces where the similarity measures cannot be evaluated. C. (top) Spike trains generated by SAP 10 with a weight of 550 pA, with a magnification for improved visualization of single burst of spikes (bottom).

for each F, $n = 21$ spike train pairs were generated. In Fig. 9C (left) two spike train pairs with high ($F = 0$) and low ($F = 1$) synchrony are shown.

3. Results

3.1. Rationale for the proposed method

A current limitation observed in most established spike train similarity measures is their inability to accurately quantify the similarity between pairs of spike trains in certain key scenarios as depicted in Fig. 7. To this aim, we compare the similarity measures (18) with *Spike-contrast*, *SPIKE-distance*, *A-SPIKE-distance*, *SPIKE-synchronization*, and *A-SPIKE-synchronization* by using four ad hoc generated spike train pairs. We start by analyzing the spike train pairs reported in panel A and B of Fig. 7. In this scenario, since *ST-Precision*, *ST-Recall*, and *ST-Fscore* are null in both cases, the method accurately detects that the spike train pairs are fully desynchronized. However, a low but non-zero value for the *ST-Accuracy* measure in both cases highlights the occurrence of common periods of inactivity, which may be indicative of some inhibition being imposed in both cells. Overall, the similarity measures in *STSimM* account for both fully desynchronized spike times and the shared periods of inactivity, which may be information-rich. On the contrary, *Spike-contrast*, *SPIKE-distance*, and *A-SPIKE-distance* quantify a non-zero spike synchronization in both cases. However, while *Spike-contrast* determines low values of synchrony, ranging from 0.249 in case A to 0.156 in case B, *SPIKE-distance* and *A-SPIKE-distance* even detect a slight improvement in case B (0.575) compared to case A (0.556). Then, both *SPIKE-distance* and *A-SPIKE-distance* measures appear to be more sensitive to the smaller difference in the number of spikes between the two (mostly uncorrelated) spike trains in panel B compared to that in panel A, whereas *Spike-contrast* quantifies the similarity in the opposite manner. Differently, *SPIKE-synchronization*

and its adaptive variant *A-SPIKE-synchronization* correctly detect the desynchronization in both spike train pairs.

The spike trains at the top of both panels C and D are identical, whereas the spike trains at the bottom of these panels differ only in the last spike, which is shifted by about 20 ms. In this case, the *ST-Accuracy* decreases from 0.558 in case C to 0.527 in case D, while the other measures in *STSimM* are lower but non-zero. Despite the shared silent period of about 600 ms within 800 ms, the low values of the *ST-Accuracy* (below 0.56) account for 10 and 11 mismatched spikes (FP) between the two pairs of spike trains in cases C and D, respectively. Furthermore, the percentage of relevant spike times detected (*ST-Precision*) is 0.577 in case C and 0.410 in case D, while the fraction of correctly detected spike times (*ST-Recall*) is lower in both cases (0.449 in case C and 0.410 in case D). The lower values of the *ST-Fscore* measure allow to highlight the detection of an high number of both FP and FN (0.313 in case C and 0.250 in case D). On the other hand, in contrast to *SPIKE-distance*, the similarity measures *SPIKE-synchronization* and its adaptive variant, as well as *Spike-contrast*, appear to be insensitive to the shift of the last spike in the second train, quantifying the level of synchrony as 0.375 and 0.5 for both cases, respectively. Finally, *SPIKE-distance* and *A-SPIKE-distance*, unexpectedly, exhibit a higher level of synchrony, with values of 0.897 and 0.935 for case C, and 0.884 and 0.923 for case D, respectively.

All *STSimM* measures are defined in a similar manner, starting from the elements of the confusion matrix. However, only *ST-Accuracy* can be considered sensitive to silence, as it is able to emphasize both silent periods (TN) and the detection of single spike times (TP, FP, and FN). Conversely, all the other measures in *STSimM* consider only the accurate detection of relevant spike times as significant. Nevertheless, since both silent periods and single spike events contribute equally to the similarity between two or more spike trains, it is appropriate to consider these quantities simultaneously. For the sake of simplicity, in the following sections, we streamline the analysis by focusing solely on *ST-Accuracy* and *ST-Fscore*, as these two measures effectively capture the essential aspects of spike train similarity, including sensitivity to silent periods and accuracy in spike detection. Nevertheless, the *STSimM* tool provides the values of all ST measures in each run. For similar reasons, we will focus exclusively on *SPIKE-distance*, *SPIKE-synchronization*, and *A-SPIKE-synchronization* in the subsequent sections.

Furthermore, all the proposed measures contain free parameters (ω, Λ, c), hence allowing for enhanced control over the analysis. In fact, an appropriate selection of these parameters on one hand allows to highlight some relevant features present in the data, and on the other hand allows to emphasize the silent periods over spike events or vice versa. However, all the ST-measures become parameter-free when the automatic selection is set as input. In this case, it is still advisable to monitor the value of Λ calculated by the STSimM tool to ensure that it aligns with time measures that correctly define two spike trains as synchronous (see Section 2.2.1 for more details).

3.2. Comparison of the performance and similarity measures STSimM, Spike-contrast, SPIKE-distance, SPIKE-synchronization, and A-SPIKE-synchronization

In this section, we compare the similarity measures implemented in the *Spike-contrast*, *SPIKE-distance*, *SPIKE-synchronization*, *A-SPIKE-synchronization*, and *STSimM* tools using the four datasets described in Sections 2.4.1–2.4.4, which exhibit varying levels of similarity or synchrony.

Specifically, spike train datasets obtained from in vivo-like synaptic stimulations (refer to Section 2.4.1) are utilized to compare the *performance measures ST-Accuracy* and *ST-Fscore* (see Eq. (5)) of the A-GLIF model with those derived using *Spike-contrast*, *SPIKE-distance*, *SPIKE-synchronization*, and *A-SPIKE-synchronization*. Furthermore, we evaluate the *similarity measures* (18) on three synthetic datasets generated from Poisson spiking, bursting, and sub-bursting models described in Sections 2.4.2–2.4.4.

To quantitatively assess the predictive power of a neuron model with respect to experimental data, it can be useful to preliminarily verify that the two sets of data are at least statistically indistinguishable, thereby confirming that they are not completely uncorrelated (Marasco et al., 2023, 2024b). Hence, in Marasco et al. (2024b), we preliminarily verified statistically that the A-GLIF model was able to reproduce the spike trains in response to each of the 130 synaptic stimulation currents (p-value >0.05 for all pairwise Mann–Whitney U-test for median difference on spike trains with at least 10 spike events). Subsequently, we relied on the performance measures (5) to assess whether the A-GLIF model accurately detected the occurrence (or non-occurrence) of a spike within a specified time interval.

An example of raster plots for the reference spike train (train 1) and the A-GLIF model (train 2), generated by synaptic stimulation SAP 7 with a peak synaptic current of 650 pA during a 10-minute stimulation, is shown in panel A of Fig. 8. The results obtained by setting $\omega = 0.35, \Lambda = 10$ ms, $c = 3$ are presented in panel B of Fig. 8 together with those achieved using *Spike-contrast*, *SPIKE-distance*, *SPIKE-synchronization*, and *A-SPIKE-synchronization* measures.

As can be observed, the *ST-Accuracy* exhibits a similar range of values to all the other similarity measures. However, its values are significantly lower compared to those of the *SPIKE-distance* measures. In detail, the *ST-Accuracy* ranges from 0.734 to 1 with a mean value of 0.911 over the entire set of simulations (see Fig. 8, panel B, top right). Furthermore, the *ST-Precision* had a mean value of 0.852 over all simulations with at least one spike event, and was below 0.8 in only 28 out of 112 cases (see Supp. Fig. A.4), while *ST-Recall* had a mean value of 0.788 over all simulations with at least one spike, and it was slightly lower (<0.7) for 12 out of 112 cases (see Supp. Fig. A.4). The minimum *ST-Precision* and *ST-Recall* values (*ST-Precision*=0.333, *ST-Recall*=0.2) occurred in the extreme case (SAP 5, weight 150 pA) where the A-GLIF model generated 3 spikes for an experimental trace with 5 spikes, resulting in only 1 TP. Finally, in the simulations where the *ST-Fscore* can be computed, the measure had a mean value of 0.812, with only 10 out of 112 cases having a value just below 0.7 (see Fig. 8, panel B, bottom right). As *ST-Precision* and *ST-Recall*, the minimum value of *ST-Fscore* (*ST-Fscore*=0.250) occurred in the case of SAP 5 with a weight of 150 pA. In this particular case, both *SPIKE-synchronization* and *A-SPIKE-synchronization* take their lowest value of 0.5, while *SPIKE-distance* and *Spike-contrast* show values of 0.711 and 0.656, respectively.

Overall, the similarity measure *Spike-contrast* ranges from 0.4 to 0.975, with a mean value of 0.868 over all simulations with at least two spikes in a train (see Fig. 8, panel B, top left) whereas *SPIKE-distance* ranges from 0.698 to 1, with a mean value of 0.987 (see Fig. 8, panel B, top middle). Finally, both the *SPIKE-synchronization* and *A-SPIKE-synchronization* measures range from 0.5 to 1, with mean values of 0.856 and 0.876, respectively (see Fig. 8, panel B, bottom left and middle). We note that, overall, there are no significant differences between the values detected by the *SPIKE-synchronization* and *A-SPIKE-synchronization* tools.

We emphasize that *ST-Accuracy* evaluates the model's ability to reproduce both spiking and silent periods in each experimental trace. In contrast, all other ST-performance measures (see Eq. (5)) do not explicitly consider silent periods, except for mismatched spike events within these periods (FP). To better illustrate this distinction in the quantitative assessment of the A-GLIF model, in panel C of Fig. 8, we display the raster plots of the spike train pair generated by SAP 10 with a weight of 550 pA (marked by a square in all heat maps in panel B of Fig. 8). This case refers to the lower value of *ST-Accuracy*, which is 0.734, corresponding to a drastic change in the firing regime (see Figs. 10 and A2 in Marasco et al. (2024b)). Furthermore, in this particular case, our performance measures exhibit significantly lower values compared to *Spike-contrast* (0.947) and *SPIKE-distance* (0.996). On the contrary, both *SPIKE-synchronization* and *A-SPIKE-synchronization* detect a lower value of 0.661, which lies between the values of *ST-Accuracy* (0.734) and *ST-Fscore* (0.589). The overall comparison of the two spike trains in panel

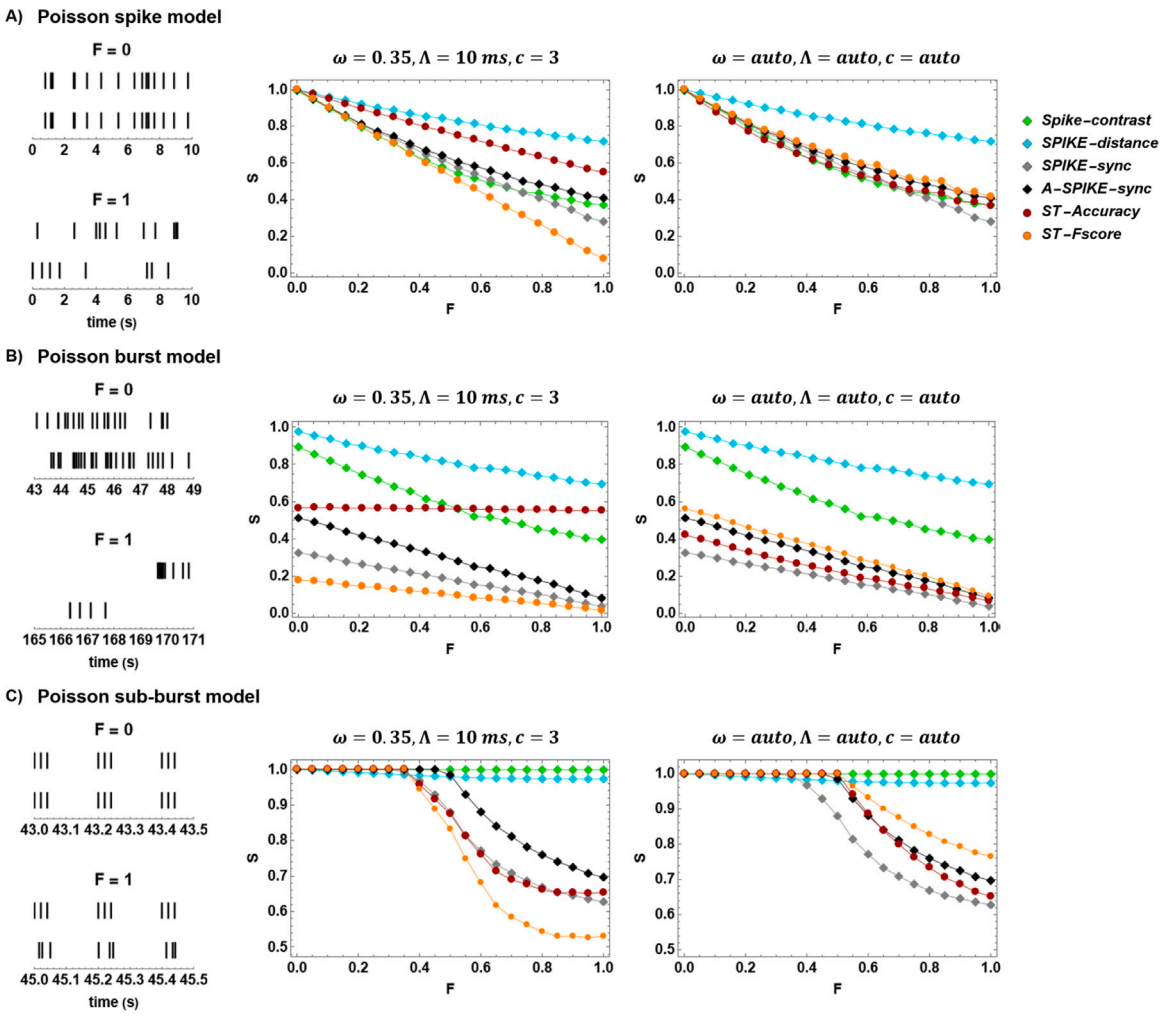


Fig. 9. Comparison of *Spike-contrast*, *SPIKE-distance*, *SPIKE-synchronization*, and *A-SPIKE-synchronization* with *STSimM* using Poisson spike (A), Poisson burst (B) and Poisson sub-burst (C) test datasets. Left: Magnification of test datasets with high ($F = 0$) and low ($F = 1$) synchrony. Middle: Mean synchrony values (S) w.r.t. F of *Spike-contrast*, *SPIKE-distance*, *SPIKE-synchronization*, *A-SPIKE-synchronization* and *STSimM*. Each data point represents the mean synchrony value S of $n = 20$ spike train pairs obtained for each synchrony level F ($F = 0, 0.05, 0.1, \ldots, 1$). Right: Same as the middle panel but with automatic parameter selection for ST-measures.

C at the top seems to confirm the perfect similarity between the spike trains as detected by *Spike-contrast* and *SPIKE-distance*. Effectively, during the long synaptic stimulation of 600 s, macroscopically the spike trains seem to share both the silent periods and the "burst activities". However, when we look at the magnification of a single interval in which a "burst activity" seems to occur, we realize that on a scale of 5 s, what appears as a typical "burst activity" indeed seems to be an interval where regular spiking activity is interleaved with some silent periods. Then, when we change the time scale, it becomes evident that the detection of the spike times is not correct as one would expect. This completely justifies the low values computed by *STSimM* for the ST-Precision (0.5), ST-Recall (0.717), and F-score (0.589), whereas a higher value for the ST-Accuracy is linked to the presence of long silent intervals. Qualitatively, it appears that the *Spike-contrast* and *SPIKE-distance* measures are sensitive to bursts, while neglecting smaller time scales. From this perspective, *STSimM* outperforms *Spike-contrast* and *SPIKE-distance*. In contrast, *SPIKE-synchronization*, and *A-SPIKE-synchronization* both appear to be more sensitive to small scales, and the common value detected by these measures aligns more closely with those obtained from the ST performance measures. Finally, we note that for these in vivo-like synaptic stimulations, the results obtained using the automatically determined set of parameters ω, Λ, and c can lead to inaccurate results, as in these cases, due to the presence of long silent intervals, the values of Λ determined by Eq. (16) are generally too high to evaluate model performance (see Supp. Fig. A.3, left panel). In Fig. 9, following Ciba et al. (2018), we compare the similarity measures (18) for the three spike train datasets with different synchrony levels determined by the factor F, ranging from 0 (highest synchrony level) to 1 (lowest synchrony level) with those derived from *Spike-contrast*, *SPIKE-distance*, *SPIKE-synchronization*, and *A-SPIKE-synchronization*. On the left of all panels in Fig. 9 we depict magnifications of typical spike train pairs, highlighting that only for the Poisson spike and Poisson sub-burst models with $F = 0$ both spike trains were identical. Conversely, for the Poisson burst model, only the burst center points of both spike trains were synchronized, but not the spike times within a burst. A statistical analysis confirms that the spike train pairs in all three datasets are statistically distinguishable in a significant number of cases (see Suppl. Fig. A.1). This is particularly the case for the Poisson burst model, in which 242 out of 400 spike train pairs are statistically distinguishable. Although in this case it may not be appropriate to conduct an analysis to quantify the similarity levels of pairs of spike trains, we proceed nevertheless to facilitate the comparison of *STSimM* with respect to *Spike-contrast*, *SPIKE-distance*, *SPIKE-synchronization*, and *A-SPIKE-synchronization*. For the same reasons, numerical results on similarity measures will be provided not for individual pairs as in the case of synaptic inputs, but as averages of the values obtained for the 20 pairs of spike trains corresponding to each value of F in each of the three datasets. Consequently, comparisons between the

measures in *STSimM* with respect to *Spike-contrast*, *SPIKE-distance*, *SPIKE-synchronization*, and *A-SPIKE-synchronization* will only provide an average level of similarity for each value of F in each of the three datasets, both when $\omega = 0.35$, $\Lambda = 10$ ms, and $c = 3$ (Fig. 9, middle panels), and for the automatically determined set of these parameters (Fig. 9, right panels).

As expected, all the similarity measures assume a value of 1 for $F = 0$, and it decrease as a function of F for $F > 0$ in the Poisson spike and Poisson sub-burst models (Fig. 9A, C), except for *SPIKE-distance* and *Spike-contrast*, which remain almost constant in the Poisson sub-burst model (see Fig. 9 C). Moreover, for the Poisson spike model, when $\omega = 0.35$, $\Lambda = 10$ ms, and $c = 3$ (see Fig. 9A, middle panel), the *ST-Accuracy* curve lies below the curve corresponding to the mean similarity measure of *SPIKE-distance* and above all the others, but for all the other similarity measures (18) the curves overlap and assume the lowest values (Fig. 9A, middle panel). In this case, the similarity measures *ST-Precision*, *ST-Recall*, and *F-score* appear to be more sensitive to precise detection of spike times, while *ST-accuracy* maintains higher values depending on shared silent periods. Nevertheless, with automatic parameter selection, the *ST-Accuracy* curve remains below the *SPIKE-distance* curve, but it overlaps with the *Spike-contrast* curve (see Fig. 9A, right panel). In this case, all similarity measures, except for *SPIKE-distance*, are comparable, with the *ST-Fscore* curve overlapping the *A-SPIKE-synchronization* curve. The higher *ST-Fscore* values observed with the automatic selection of parameters ω, Λ, and c, compared to values $\omega = 0.35$, $\Lambda = 10$ ms, and $c = 3$, are attributed to the *STSimM* tool selecting Λ values between 200 ms and 300 ms (see Supp. Fig. A.3), which allows for less precision in spike detection.

Furthermore, for the Poisson burst model, the similarity measures (18), *SPIKE-synchronization*, and *A-SPIKE-synchronization* significantly differ from both *Spike-contrast* and *SPIKE-distance* (see Fig. 9B). In this case, *Spike-contrast* and *SPIKE-distance* show significantly higher average similarity values compared to all other measures, even though for this dataset the majority of spike train pairs were statistically distinguishable. Conversely, when $\omega = 0.35$, $\Lambda = 10$ ms, and $c = 3$ our similarity measures *ST-Precision*, *ST-Recall*, and *ST-Fscore* dramatically decrease from 0.18 to 0.01 because the spikes in all sub-bursts are desynchronized, whereas *ST-Accuracy* shows a quasi-constant behavior around a value of 0.55, as this measure is affected by silent periods. For this dataset, *SPIKE-synchronization* and *A-SPIKE-synchronization* reached a maximum value of 0.32 and 0.51, respectively, when $F = 0$. and decrease w.r.t. F, converging to 0 as F approaches 1, since the spike train pairs become fully desynchronized when $F = 1$. We would emphasize that, with the automatic selection of the parameters, the similarity measures (18) all decrease with respect to F (see Fig. 9B, right panel). Specifically, the *ST-Accuracy* curve lies below that of all other ST measures and between those of *A-SPIKE-synchronization* and *SPIKE-synchronization*. Additionally, the *ST-Fscore* curve exhibits behavior similar to *A-SPIKE-synchronization*, as the values of Λ, which represent the maximum half-width of each interval in which two spike times are considered synchronous, range from 1.8 to 2.8 s (see Supp. Fig. A.3).

Finally, for the Poisson sub-burst model, *Spike-contrast* and *SPIKE-distance* maintain an almost constant value around 1 for all F, whereas *SPIKE-synchronization* and *A-SPIKE-synchronization* reach this value only for $F \leq 0.3$ and $F \leq 0.45$, respectively (see Fig. 9C). Similarly, all similarity measures (18) assume a value of 1 only for $F \leq 0.35$ and $F \leq 0.5$, when $\omega = 0.35$, $\Lambda = 10$ ms, and $c = 3$ or when the automatic parameter selection procedure is implemented, respectively. (Fig. 9C, middle and right panels). Moreover, starting from $F = 0.35$ (or $F = 0.5$) the behavior diverges, and the curves of the mean values of (18), *SPIKE-synchronization*, and *A-SPIKE-synchronization* rapidly decrease. For this dataset, all similarity measures (18), *SPIKE-synchronization*, and *A-SPIKE-synchronization* appear to be more sensitive to both the precise detection of spike times and to shared silent periods. In this case, the selection of parameters ω, Λ, and $c = 3$ results in markedly different behaviors in the *ST-Accuracy* and *ST-Fscore* curves. Specifically, with $\omega = 0.35$, $\Lambda = 10$ ms, and $c = 3$, the ST measures exhibit lower values compared to those obtained with automatic parameter selection. In this configuration, *ST-Accuracy* aligns with the *SPIKE-synchronization* curve, whereas with automatic parameter selection, it aligns with the *A-SPIKE-synchronization* curve. Notably, the lowest values for *ST-Fscore* are associated with the parameters $\omega = 0.35$, $\Lambda = 10$ ms, and $c = 3$, since the automatic selection yields nearly constant Λ values around 131 ms for this dataset (see Supp. Fig. A.3). As a result, less precision is required in determining when two spikes are considered synchronous.

We emphasize that the proposed measures exhibit a strong adaptive nature, enabling customization according to the intrinsic characteristics of the data. Specifically, these measures offer the flexibility to adjust to different patterns of neuronal activity, such as precise spike timing in both regular spiking and bursting, as well as the detection of silent periods. This adaptability is particularly advantageous under specific experimental conditions, as it enhances sensitivity to subtle variations within the data. Such flexibility renders our approach especially effective for a wide range of neuroscientific analyses.

3.3. Statistical analysis of datasets and correlation of the similarity measures

The overall results presented in Figs. 7–9 demonstrate that *STSimM* is more sensitive than *Spike-contrast* and *SPIKE-distance* in detecting the level of similarity of spike train pairs. In fact, although both *Spike-contrast* and *SPIKE-distance* measures are independent of time scales, i.e., they automatically detect the time scales within spike train data, they appear to be insensitive to small time scales when analyzing datasets containing multiple time scales. Moreover, the results obtained with the *STSimM* tool are overall more aligned with those of the *SPIKE-synchronization* and *A-SPIKE-synchronization* measures, particularly in cases of highly desynchronized spike trains. However, automatic timescale detection can reduce synchronization precision when long silent intervals are present. The *STSimM* tool addresses this by displaying the Λ value, representing the maximum half-width of each interval in which two spikes are considered synchronous, allowing users to select a Λ value better suited to their specific dataset.

Furthermore, we observe that only the spike train pairs of the Poisson sub-burst model were statistically indistinguishable in all cases (see Suppl. Fig. A.1, right panel for the Mann–Whitney U-test results at the 0.05 significance level), as the majority of the 400 spike train pairs from the Poisson spike model (353 out of 400). In contrast, only 158 out of 400 spike train pairs from the Poisson sub-burst model were statistically indistinguishable (see Suppl. Fig. A.1, left and middle panels). In this last case, evaluating the level of similarity appears to have limited significance, except to confirm the reasons for the low levels of similarity detected by *STSimM*, *SPIKE-synchronization* and *A-SPIKE-synchronization* tools. Nevertheless, when evaluating the *Spearman correlation* among all the similarity measure, as in Ciba et al. (2018), we observed correlations for all datasets, except for those obtained from in vivo-like synaptic stimulations (see Supp. Table A.1 and Suppl. Fig. A.2). Specifically, in this case, the ST measures correlate only with *SPIKE-synchronization* and *A-SPIKE-synchronization*. This result should not be surprising, as the spike train pairs considered originate from in vivo-like synaptic stimulations of a biophysically accurate neuron model, unlike Poisson spike, burst, and sub-burst models (Marasco et al., 2024b). Consequently, their structure is not affected by the statistical manipulation employed to generate the Poisson models.

3.4. Comparison of the computational times

To accelerate computational time and fully utilize the multiprocessing module for evaluating the similarity measures in *STSimM*, *SPIKE-distance*, *Spike-contrast*, *SPIKE-synchronization*, and *A-SPIKE-*

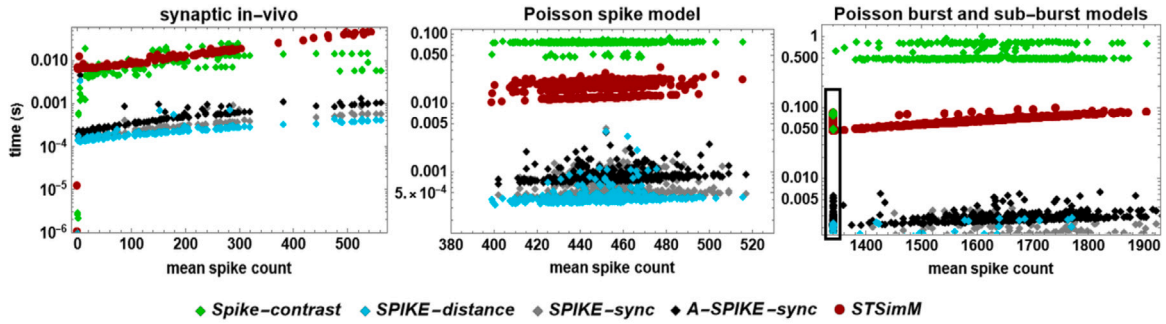


Fig. 10. Computational speed. A. Computational time required for *Spike-contrast* (green squares), *SPIKE-distance* (cyan squares), *SPIKE-synchronization* (gray diamonds), *A-SPIKE-synchronization* (black diamonds), and *STSimM* (red circles) similarity measures as a function of the average number of spikes in spike train pairs across synaptic in-vivo dataset (left), Poisson spike (middle), Poisson sub-burst (highlighted in the black box), and Poisson burst models (right). In all cases, except for synaptic in-vivo data where performance measures are evaluated with $\omega = 0.35$, $\Lambda = 10$ ms, and $c = 3$, the computational time of the *STSimM* tool for the similarity measures corresponds to that of the automatic parameter selection.

synchronization, we employed a login node with 2 CPUs of an HPC cluster from the EBRAINS-Italy infrastructure (https://www.ebrains-italy.eu). The computational efficiency of our performance and similarity measures have been further improved through the cythonization of the Python script. In Fig. 10, we compare the computational time per mean spike count for *Spike-contrast*, *SPIKE-distance*, *SPIKE-synchronization*, *A-SPIKE-synchronization*, and *STSimM* using the datasets described in Section 2.4. We observe that for the Poisson spike and burst datasets, our method is computationally more efficient than *Spike-contrast* , while for the synaptic in vivo, and Poisson sub-burst datasets (with a fixed mean spike count of 1341 across the Poisson dataset), our method has comparable computational time. Our method generally performs better computationally on shorter spike trains (e.g., 300 s as in the case of Poisson spike and sub-burst models) and worse on longer spike trains (e.g., 3600 s as in the case of Poisson burst model). Overall, the *STSimM* tool is considerably less computationally efficient than *SPIKE-distance*, *SPIKE-synchronization*, and *A-SPIKE-synchronization*. However, it provides all four performance measures — *ST-Accuracy*, *ST-Precision*, *ST-Recall*, and *ST-Fscore* — in each run.

4. Discussion

Measures to identify the similarity among spike trains are applied in two major scenarios: to assess the quality of predictions of (spiking) neuron models, and to estimate the degree of similarity between two or more spike trains. In the literature, there are at least 34 correlation or similarity measures (see Cutts and Eglen, 2014 for a detailed comparison). However, most of them are sensitive to user-defined fixed time scales, thus ignoring the different time scales typically present in real-neuron data, ranging from regular spiking to bursting, and often interleaved with silent periods. The most popular time scale-dependent measures used are the Victor and Purpura (1996) and van Rossum (2001) distances. Complementary approaches that are adaptive to time scales and often free of parameters have been proposed (see Cutts and Eglen, 2014; Kreuz et al., 2015; Satuvuori et al., 2017; Ciba et al., 2018). Although the synchrony measures embedded in *Spike-contrast* and *SPIKE-distance* are independent of time scales, they appear to overlook the importance of accurately detecting the timing of individual spikes (see Figs. 7C–D, 8B–C, 9B–C).

Conversely, *SPIKE-synchronization* and *A-SPIKE-synchronization* appear to be more sensitive to fully or partially desynchronized spike trains (see Figs. 7A–B, 9B–C). However, their automatic timescale detection can reduce synchronization precision when long silent intervals are present (see Figs. 5, 9C), and generally disregards periods of inactivity (Figs. 7A–B, 8B). In contrast, all measures of the *STSimM* tool, particularly *ST-Fscore*, ensure precise identification of spike occurrences using the parameter $\omega \leq 0.5$, with a maximum shift of $\pm\Lambda$ ms that can be defined by the user (see Figs. 2, 4, and 5). In particular, a maximum shift of ±10 ms ensures precise identification of spike occurrences for CA1 neuron in-vivo activities (Marasco et al., 2024b), whereas other studies suggest detecting a spike event with a precision of less than ±5 ms (Kara et al., 2000; Reinagel and Reid, 2002).

Another important aspect of spike train analysis involves shared inactivity or silent periods (Lyttle and Fellous, 2011), while most similarity measures appear to overlook this aspect or even exclude it when computing correlations (Cutts and Eglen, 2014). In particular, *Spike-contrast* requires that at least one of the two spike trains contains two or more spikes, while *ST-Fscore* can only be determined if at least one of the two spike trains includes at least one spike. Conversely, *ST-Accuracy*, *SPIKE-distance*, *SPIKE-synchronization*, and their adaptive variants can always be evaluated, even when both spike trains are empty. In our framework, the *ST-Accuracy* measure, besides accounting for precise spike time detection, can identify shared silent periods of inactivity between spike trains. To account for this characteristic, it is sufficient to assign an appropriate value to the parameter c, which controls the extent to which silent periods are emphasized.

Despite the computational efficiency of our performance and similarity measures have been improved through the utilization of the Python multiprocessing module and the cythonization of the Python script, the *STSimM* tool remains significantly less computationally efficient than *SPIKE-distance*, *SPIKE-synchronization*, and *A-SPIKE-synchronization*. Nevertheless, for the Poisson spike and burst datasets, our method exhibits superior computational performance compared to *Spike-contrast*. However, for the synaptic in vivo, and Poisson sub-burst datasets, has comparable computational time to *Spike-contrast*. Moreover, *STSimM* tool remains computationally intensive when computing measures for more than two spike trains, indicating the potential for further optimization through parallelization techniques.

Although we observed a high correlation between *STSimM* and all other synchrony measures for most datasets, our measures appear to be better suited for detecting both the precise timing of single spikes and shared silent periods. Notably, for the synaptic in-vivo spike dataset, our measures correlate only with *SPIKE-synchronization* and *A-SPIKE-synchronization*, demonstrating a closer alignment of the *STSimM* tool with these measures in accurately detecting spike times. Furthermore, the inclusion of three key measures allows for an easy distinction of similarity levels across neuronal activity, whether interleaved with silent periods (*ST-Accuracy)* or solely focusing on spike timing accuracy (*ST-Precision*, *ST-Recall*, and *ST-Fscore*). Furthermore, the integration of three free model parameters that allow both precise spike detection (ω, Λ) and the weighting of silent periods (c) provides users with

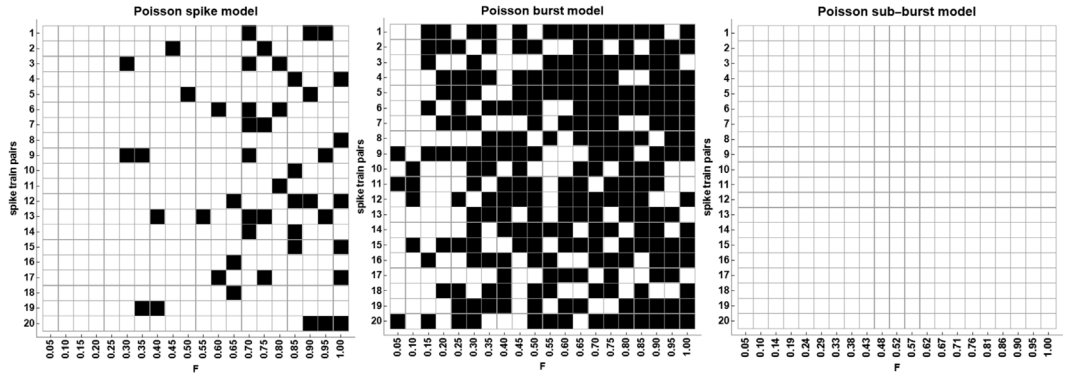


Fig. A.1. Statistical indistinguishability of spike train datasets. Graphical representation of the p-values from the Mann–Whitney U-test for spike train pairs in the three datasets. Black and white squares represent p-values less than or greater than 0.05, respectively.

Table A.1
Spearman correlation coefficients. Lack of correlation is indicated in bold.

	Spike-contrast	SPIKE-distance	SPIKE-sync	A-SPIKE-sync	ST-Accuracy	ST-Precision	ST-Recall	ST-Fscore	Spike-contrast	SPIKE-distance	SPIKE-sync	A-SPIKE-sync	ST-Accuracy	ST-Precision	ST-Recall	ST-Fscore
Spike-contrast	1.	**0.382**	**0.231**	**0.237**	**0.116**	**0.017**	**0.197**	**0.254**	1.	0.996	0.996	0.996	0.995	0.996	0.996	0.996
SPIKE-distance		1.	**0.61**	**0.665**	**0.448**	**0.349**	**0.465**	**0.637**		1.	0.996	0.996	0.996	0.996	0.996	0.996
SPIKE-sync			1.	0.937	0.885	**0.621**	0.801	0.959			1.	0.998	0.998	0.998	0.998	0.998
A-SPIKE-sync				1.	0.801	**0.567**	0.78	0.933				1.	0.996	0.996	0.996	0.996
ST-Accuracy					1.	0.845	**0.685**	0.912					1.	1.	1.	1.
ST-Precision						1.	**0.299**	**0.667**						1.	1.	1.
ST-Recall							1.	0.823							1.	1.
ST-Fscore								1.								1.
	synaptic in-vivo											**Poisson**				
Spike-contrast	1.	0.992	0.989	0.991	0.854	0.987	0.988	0.988	1.	0.877	0.861	0.805	0.85	0.85	0.85	0.85
SPIKE-distance		1.	0.99	0.991	0.858	0.987	0.988	0.987		1.	0.971	0.931	0.969	0.968	0.968	0.968
SPIKE-sync			1.	0.998	0.877	0.995	0.996	0.996			1.	0.961	0.983	0.984	0.984	0.984
A-SPIKE-sync				1.	0.867	0.995	0.995	0.995				1.	0.962	0.963	0.963	0.963
ST-Accuracy					1.	0.895	0.892	0.894					1.	1.	1.	1.
ST-Precision						1.	1.	1.						1.	1.	1.
ST-Recall							1.	1.							1.	1.
ST-Fscore								1.								1.
	Poisson Burst											**Poisson Sub-Burst**				

additional flexibility in spike train analysis. Finally, we found that the quantitative results reported in the previous section were unaffected by the numerical choice of the parameters ω, Λ, and c provided that the numerical values used were reasonably physiological (results not shown). Even though originally developed within a neuroscientific framework to assess the performance of spiking neuron models and the similarity between pairs of spike trains, the measures implemented in *STSimM* can be applied to general discrete time series consisting of sequences of events of any nature. To this end, an automatic procedure is also available to set the numerical values of the three parameters, ω, Λ, and c.

CRediT authorship contribution statement

A. Marasco: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **C.A. Lupascu:** Writing – review & editing, Writing – original draft, Validation, Software, Investigation. **C. Tribuzi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation.

Funding

This paper has been funded by the Italian National Recovery and Resilience Plan (NRRP), M4C2, funded by the European Union – NextGenerationEU (Project IR0000011, CUP B51E22000150006, 'EBRAINS-Italy'), and by the EU Horizon Europe Program under the specific Grant Agreement 101147319, EBRAINS 2.0 project.

Declaration of competing interest

All the authors declare that no competing interests exist.

Acknowledgments

We thank Alessia Bonafede for technical and administrative assistance. A.M. acknowledges membership of the INdAM-GNFM Group.

Appendix. Supporting information

See Table A.1 and Figs. A.1–A.4.

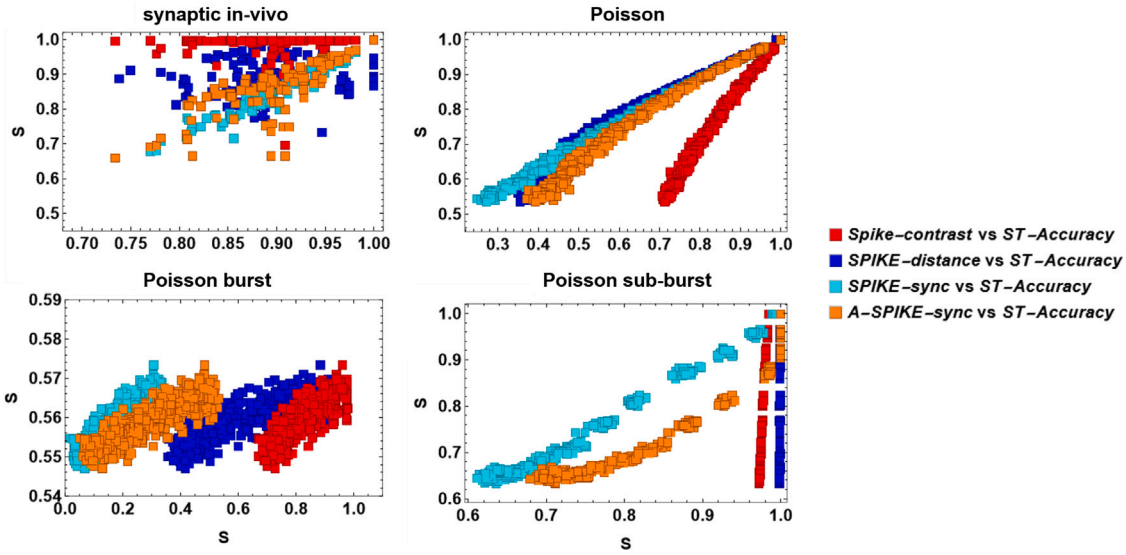


Fig. A.2. Synchrony values of *Spike-contrast* versus *ST-Accuracy* (red squares), SPIKE-distance versus *ST-Accuracy* (blue squares), *SPIKE-synchronization* versus *ST-Accuracy* (cyan squares), and *A-SPIKE-synchronization* versus *ST-Accuracy* (orange squares) for all spike train datasets.

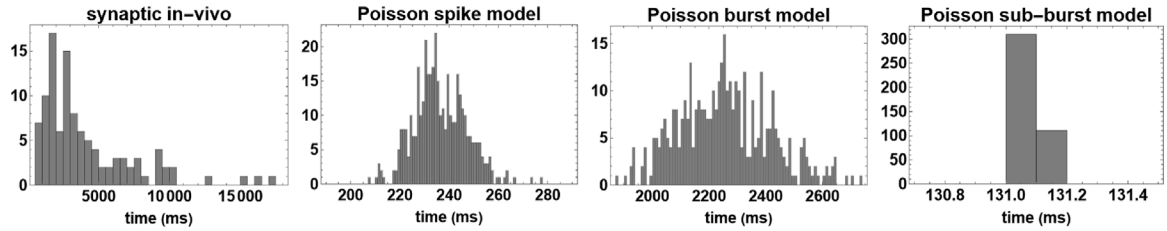


Fig. A.3. Automatic values of Λ obtained from Eq. (16) for each dataset.

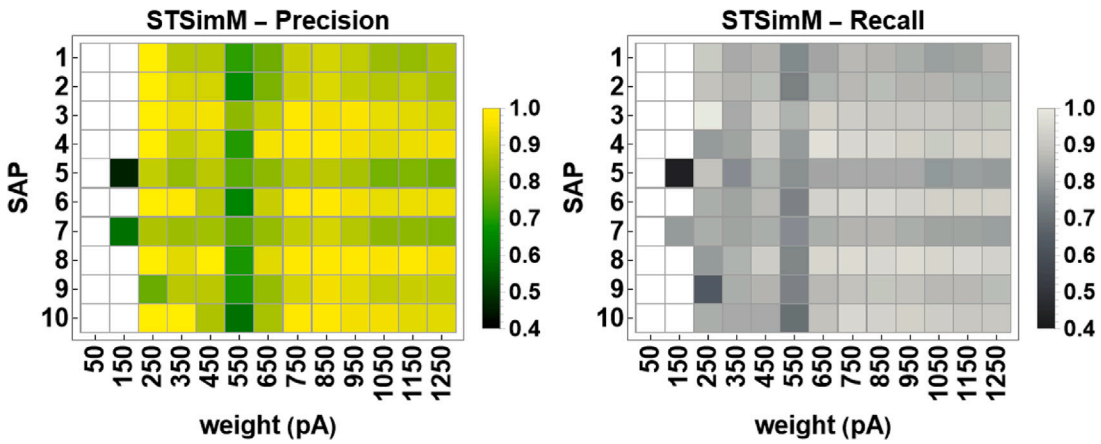


Fig. A.4. Synaptic in-vivo like stimulations. *ST-Precision* and *ST-Recall* performance measures (5) for the references and A-GLIF model traces obtained in response to 130 synaptic stimulations.

Data availability

All Python scripts and datasets are available for download from the ModelDB section of the Senselab database (https://modeldb.science/2016671) and at https://github.com/biomemsLAB/Spike-Contrast/blob/master/Testdata.zip.

References

Ciba, M., Isomura, T., Jimbo, Y., Bahmer, A., Thielemann, C., 2018. Spike-contrast: A novel time scale independent and multivariate measure of spike train synchrony. J. Neurosci. Methods 293, 136–143. http://dx.doi.org/10.1016/j.jneumeth.2017.09.008.

Cutts, C.S., Eglen, S.J., 2014. Detecting pairwise correlations in spike trains: An objective comparison of methods and application to the study of retinal waves. J. Neurosci. 34 (43), 14288–14303. http://dx.doi.org/10.1523/JNEUROSCI.2767-14.2014.

Kara, P., Reinagel, P., Reid, R.C., 2000. Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. Neuron 27 (3), 635–646. http://dx.doi.org/10.1016/S0896-6273(00)00072-6.

Kreuz, T., Chicharro, D., Greschner, M., Andrzejak, R.G., 2011. Time-resolved and time-scale adaptive measures of spike train synchrony. J. Neurosci. Methods 195 (1), 92–106. http://dx.doi.org/10.1016/j.jneumeth.2010.11.020.

Kreuz, T., Chicharro, D., Houghton, C., Andrzejak, R.G., Mormann, F., 2013. Monitoring spike train synchrony. J. Neurophysiol. 109 (5), 1457–1472. http://dx.doi.org/10.1152/jn.00873.2012.

Kreuz, T., Mulansky, M., Bozanic, N., 2015. SPIKY: a graphical user interface for monitoring spike train synchrony. J. Neurophysiol. 113 (9), 3431–3445. http://dx.doi.org/10.1152/jn.00848.2014.

Lyttle, D., Fellous, J., 2011. A new similarity measure for spike trains: sensitivity to bursts and periods of inhibition. J. Neurosci. Methods 199 (2), 296–309. http://dx.doi.org/10.1016/j.jneumeth.2011.05.005.

Marasco, A., Spera, E., Falco, V.D., Iuorio, A., Lupascu, C., Solinas, S., Migliore, M., 2023. An adaptive generalized leaky integrate-and-fire model for hippocampal CA1 pyramidal neurons and interneurons. Bull. Math. Biol. 85 (11), 109. http://dx.doi.org/10.1007/s11538-023-01206-8.

Marasco, A., Tribuzi, C., Iuorio, A., Migliore, M., 2024a. Mathematical generation of data-driven hippocampal CA1 pyramidal neurons and interneurons copies via A-GLIF models for large-scale networks covering the experimental variability range. Math. Biosci. 371, 109179. http://dx.doi.org/10.1016/j.mbs.2024.109179, URL: https://www.sciencedirect.com/science/article/pii/S0025556424000397.

Marasco, A., Tribuzi, C., Lupascu, C., Migliore, M., 2024b. Modeling realistic synaptic inputs of CA1 hippocampal pyramidal neurons and interneurons via adaptive generalized leaky integrate-and-fire models. Math. Biosci. 372, 109192. http://dx.doi.org/10.1016/j.mbs.2024.109192.

Mulansky, M., Bozanic, N., Sburlea, A., Kreuz, T., 2015. A guide to time-resolved and parameter-free measures of spike train synchrony. In: 2015 International Conference on Event-Based Control, Communication, and Signal Processing. EBCCSP, pp. 1–8. http://dx.doi.org/10.1109/EBCCSP.2015.7300693.

Reinagel, P., Reid, R.C., 2002. Precise firing events are conserved across neurons. J. Neurosci. 22 (16), 6837–6841. http://dx.doi.org/10.1523/JNEUROSCI.22-16-06837.2002.

van Rossum, M., 2001. A novel spike distance. Neural Comput. 13 (4), 751–763. http://dx.doi.org/10.1162/089976601300014321.

Satuvuori, E., Mulansky, M., Bozanic, N., Malvestio, I., Zeldenrust, F., Lenk, K., Kreuz, T., 2017. Measures of spike train synchrony for data with multiple time scales. J. Neurosci. Methods 287, 25–38. http://dx.doi.org/10.1016/j.jneumeth.2017.05.028.

Victor, J.D., Purpura, K.P., 1996. Nature and precision of temporal coding in visual cortex: a metric-space analysis. J. Neurophysiol. 76 (2), 1310–1326. http://dx.doi.org/10.1152/jn.1996.76.2.1310.