Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Optimal Transport Theory to Extract Spiking Motifs

Antoine Grimaldi

antoine.grimaldi@cnrs.fr

Brain and Cognition Research Center

Boris Sotomayor-Gómez

Center for Neurology, Department of Parkinson, Sleep and Movement Disorders

Matthieu Gilson

Institut de Neurosciences de la Timone

Laurent Perrinet

Institut de Neurosciences de la Timone

Andrea Alamia

Brain and Cognition Research Center

Martin Vinck

Ernst Strüngmann Institute for Neuroscience

Research Article

Keywords: Optimal Transport Theory, Temporal Codes, Spiking Motifs, Neural Decoding

Posted Date: April 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9518938/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Optimal Transport Theory to Extract Spiking Motifs

Antoine Grimaldi[1*], Boris Sotomayor-Gómez[2,3,4], Mathieu Gilson[5], Laurent U Perrinet[5], Andrea Alamia[1†], Martin Vinck[2,6†]

[1]Centre de Recherche Cerveau et Cognition (CerCo), UMR 5549, CNRS, Université Toulouse III – Paul Sabatier, Toulouse, France.
[2]Ernst Strüngmann Institute for Neuroscience, Ernst Strüngmann Institute, Frankfurt am Main, Germany.
[3]Center for Neurology, Department of Parkinson, Sleep and Movement Disorders (DZNE), Bonn, Germany.
[4]Institute of Informatics, Austral University of Chile, Valdivia, Chile.
[5]Institut de Neurosciences de la Timone (INT), UMR 7289, CNRS, Aix-Marseille Université, Marseille, France.
[6]Donders Centre for Neuroscience, Radboud University Nijmegen, Nijmegen, Netherlands.

*Corresponding author(s). E-mail(s): antoine.grimaldi@cnrs.fr;
†These authors contributed equally to this work.

Abstract

Spatiotemporal spike motifs, i.e. repeating temporal sequences, of neural activity are thought to carry information beyond average firing rates, yet extracting such motifs from noisy spike trains remains challenging. In particular, most approaches rely on bin-by-bin comparisons that are sensitive to bin size and temporal jitter. Here, we investigate the use of the Earth Mover's Distance (EMD), a metric from optimal transport theory, as a reconstruction loss for learning spiking motifs. We compare EMD to the conventional mean squared error (MSE) by training single-layer autoencoders with tied positive weights to reconstruct neural raster plots. Using synthetic spike trains with known ground-truth motifs, we systematically evaluate how each loss recovers spatiotemporal structure under varying numbers of training samples and different noise regimes, including temporal jitter, sequence warping, neuron dropout, and additive Poisson noise. We show that EMD-based training is more robust to temporal jitter and sequence warping, and more reliably estimates aligned spike timings when the number of training

samples is limited. In contrast, MSE better captures the full spike distribution, i.e. the relative timing of the spikes as well as their temporal precision, when large datasets are available and is more resilient to additive noise. Applying the same framework to Neuropixels recordings from the Allen Institute visual coding dataset reveals that motifs learned with EMD are more discriminative of visual stimuli than those learned with MSE. Together, these results highlight the advantages of optimal transport–based losses for learning spatiotemporal motifs in neural population activity.

Keywords: Optimal Transport Theory, Temporal Codes, Spiking Motifs, Neural Decoding

1 Introduction

Neural activity can be described in terms of either firing rates or precise spike timing. While rate codes capture the number of spikes within a given temporal window, temporal codes consider the exact relative timing of spikes across neurons. Both regimes likely coexist, and individual neurons may operate along a continuum between rate-based integration and temporal precision [1, 2]. Understanding how spiking motifs—reproducible sequences of spikes across multiple neurons—are represented in neural circuits is crucial to deciphering population activity beyond average firing rates.

Spiking motifs have been observed in sensory systems, where stereotypical latency patterns emerge in response to stimuli [3–5], and in cortical networks during spontaneous activity [6, 7]. These motifs capture the relative timing of spikes across neurons and provide a richer description of population activity than firing rates alone. Traditional methods for extracting motifs rely on counting repeated patterns within discretized time bins [8, 9]. While effective for certain applications, these approaches are sensitive to temporal jitter and bin size, and often scale poorly with large populations. The Earth Mover's Distance (EMD), a metric from optimal transport theory, offers a robust alternative for comparing spike trains [10]. By computing the minimal "work" required to transform one spike train into another, *Spikeship*, the metric introduced in [10], captures global temporal differences and tolerates jitter, while remaining independent of firing rates. Although Spikeship efficiently clusters spike trains by stimulus type, it does not provide a method for extracting spiking motifs in neural populations.

Here, we investigate how single-layer autoencoders (AEs) trained with different reconstruction losses—EMD versus mean squared error (MSE)—capture spiking motifs. These AEs are constrained to a single latent variable and positive, tied weights, allowing them to learn one motif per neuron population while capturing trial-to-trial variability. Such restrictions to very simple models allow us to focus on the differences induced by the reconstruction losses. By normalizing spike trains along the temporal axis, we focus exclusively on temporal structure rather than firing rate differences. We evaluate the learned motifs against ground-truth motifs in synthetic data to quantify

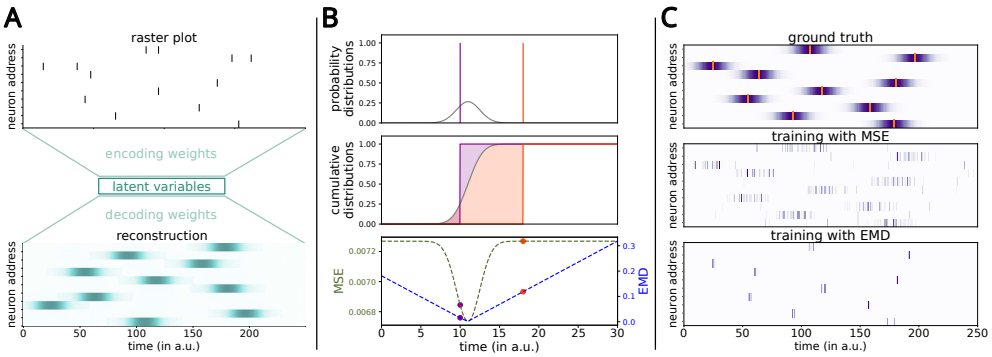


Fig. 1 Illustration of the method. **(A) Illustration of the AE applied to spike trains.** The top panel represents a synthetic raster plot, randomly generated from the structured probability distribution shown in the top figure of column C (ground truth). In general, the encoding weights transform the input into an activation in the latent space, and the input is reconstructed from these latent variables using the decoding weights. **(B) Comparison of the MSE and the EMD**. The low panel shows a comparison of the EMD and MSE values between a probability distribution (grey) and binary spikes (purple and orange), which are presented in the top panel. This illustration represents 1D signals, i.e. a single neuron address. One can see that the MSE saturates when the spike occurs outside the range of the grey probability distribution. However, since the EMD is calculated using the absolute difference between the cumulative distributions of the compared functions (see the filled areas in the middle panel), the distance linearly increases with the timing difference from the mean of the grey distribution. **(C) Example of weights trained with MSE or EMD.** The top plot represents the ground truth used to generate different samples. In purple, the probability values reflect both the relative average timings of the spikes and their precision. They follow a 1D Gaussian distribution where the average timing is the mean of the Gaussian and the precision is linked to the standard deviation of the distribution. In orange, the average timings of the spikes represent the temporal pattern without its precision. The AE is trained by minimizing the distance between the reconstruction and the input raster plot described in A. The middle plot shows the outcome of the training with 30 samples, drawn from the ground truth illustrated in the top panel, using the MSE as a loss function. One observes that the method captures the pattern in a noisy fashion by averaging, per time bin, the spikes obtained across the different samples. For the training with the EMD, the method interpolates the spike timings of the different observations. It leads to a more compact representation.

the ability of each loss to capture temporal sequences accurately and robustly, even in the presence of noise or limited training samples.

Our results show that EMD-trained AEs extract sharper, temporally aligned motifs and are more robust to temporal jitter and sequence warping. Importantly, EMD-based training performs better when the number of samples in the dataset is limited. By contrast, MSE-trained AEs are more tolerant to unstructured background spikes and can also recover the temporal precision of spikes when very large datasets are available. Applying the same framework to neural recordings from the Allen Institute for Brain Science shows that EMD-based training yields motifs that are more specific to visual stimuli than those learned with MSE. Importantly, while decoding performance based on firing rates is higher, our approach uncovers temporal motifs that provide additional, rate-independent information about population activity. Overall, our study highlights the utility of optimal transport theory for extracting spiking

motifs and systematically compares reconstruction losses in shaping learned temporal representations.

We provide a PyTorch implementation of this method and all results presented in this article can be reproduced using the code in this github repository.

2 Methods and Materials

In this study, we investigated how different reconstruction losses shape the spike motifs learned by low-dimensional generative models. We implemented a minimal single-layer AE to extract stimulus-specific spiking patterns and tested its behavior on both synthetic spike trains with controlled structure and large-scale neural recordings from the Allen Institute Visual Coding dataset [11]. The main focus is on comparing MSE and EMD as reconstruction losses, evaluating how they capture temporal structure, and robustness to noise.

2.1 Single-layer autoencoder for spike motif extraction

We consider a simple single-layer AE operating on neural raster plots. Each input sample is represented as a binary matrix $X \in \{0,1\}^{N\times T}$, where N denotes the number of neurons and T the number of time bins.

The encoder projects the input onto a one-dimensional latent variable:

$$z = W^\top X, \tag{1}$$

where $W \in \mathbb{R}_+^{N\times T}$ denotes the weight matrix. The decoder reconstructs the input as:

$$\hat{X} = zW. \tag{2}$$

Encoder and decoder weights are tied, resulting in a rank-one reconstruction of the input. Weights are constrained to be non-negative and initialized with positive values. Optimization is performed using exponentiated gradient descent, which preserves the positivity of the parameters [12]. Training is implemented in PyTorch using the Adam optimizer.

This architecture constrains the model to learn a single spike motif encoded in the weight matrix W, while the latent variable z captures trial-to-trial variability in the motif amplitude. With non-negative weights and tied parameters, the model is closely related to non-negative matrix factorization (NMF), where the weight vector corresponds to a single non-negative component [13].

Prior to training, raster plots are normalized along the temporal axis so that the total spike mass of each neuron sums to one. This removes differences in firing rates across trials and isolates temporal information.

2.2 Earth Mover's Distance as reconstruction loss

The EMD measures the minimum amount of work required to transform one probability distribution into another, where the cost corresponds to the mass moved multiplied

by the distance [14]. For one-dimensional distributions μ and ν, the EMD can be written as:

$$\mathrm{EMD}(\mu, \nu) = \int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)|\, dx, \tag{3}$$

where F_μ and F_ν are the cumulative distribution functions.

For spike trains interpreted as discrete 1D distributions along the temporal axis, the reconstruction loss is computed independently for each neuron:

$$\mathcal{L}(X, \hat{X}) = \frac{1}{N} \sum_{n=1}^{N} \mathrm{EMD}(X_n, \hat{X}_n) = \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \left| F_{X_n}(t) - F_{\hat{X}_n}(t) \right|. \tag{4}$$

Figure 1-(B) illustrates the principle of the EMD between a 1D Gaussian distribution and two Dirac distributions corresponding to binary spike events. Because the cost is computed from cumulative distributions, the EMD captures temporal discrepancies globally: the cost at a given time bin depends on differences accumulated at previous bins. In contrast, MSE evaluates errors independently in each time bin. This makes the EMD naturally sensitive to spike timing shifts, particularly in sparse spike trains.

2.3 Generative model of synthetic spike trains

To systematically evaluate reconstruction losses, we generated synthetic spike trains from a known temporal motif. Each motif specifies a set of discrete spike times $\{\mu_{n,s}\}$, with $n = 1, \ldots, N$ and s indexing spikes within each neuron.

Spike trains are generated by adding temporal jitter to each spike:

$$t_{n,s} = \mu_{n,s} + \epsilon_{n,s}, \quad \epsilon_{n,s} \sim \mathcal{N}(0, \sigma_n^2), \tag{5}$$

where σ_n controls the standard deviation of timing variability for neuron n.

Additional noise manipulations include:

- **Additive noise:** random spikes added independently across neurons and time bins.
- **Neuron dropout:** with probability p_{drop}, a neuron can be silent during some trials.
- **Temporal warping:** the sequence is compressed along the temporal axis. A random compression coefficient $C_{tw} \in [0, C_{\max}]$ is drawn for each sample. $C_{tw} = 0$ corresponds to no compression; $C_{tw} = C_{\max}$ corresponds to fully synchronized spikes.

2.4 Metrics to compare learned representations

We quantify the quality of learned spike motifs W against the ground-truth motif K (see Figure 1-(C) top panel for an example):

- **Similarity value:** cosine similarity between the flattened AE weights and the full ground-truth kernel:

$$\mathrm{Similarity} = \frac{\langle W, K \rangle}{\|W\|_2 \, \|K\|_2}. \tag{6}$$

This metric captures both spike timing and precision (temporal width of the Gaussian spike).

- **Mean timing similarity:** uses an idealized kernel with Dirac delta functions at the mean spike times of each neuron (in orange in Figure 1-(C) top panel). Cosine similarity between the AE weights and this kernel evaluates correct alignment independent of temporal precision.
- **Mean absolute timing difference:** for neurons with a single spike per motif, the estimated spike time $\hat{t}_n$ is obtained by fitting a Gaussian to the weights of neuron n. The average absolute deviation from the ground-truth spike times t_n is then computed:

$$\text{Mean absolute timing difference} = \frac{1}{N}\sum_{n=1}^{N}|\hat{t}_n - t_n|. \tag{7}$$

This metric allows a better comparison of convergence across reconstruction losses.

2.5 Allen Institute Visual Coding dataset and generative decoding

Finally, we applied our methods to neural recordings from the Allen Institute for Brain Science [11]. The dataset contains spiking activity from 32 mice presented with 20 natural images. Each image was shown 10 times per animal, and neural responses were recorded using Neuropixels probes.

Data were binned into 1 ms time bins within a 250 ms window following stimulus onset, yielding spike matrices of size $N \times 250$, where N varies across animals. Binary raster plots indicate spike occurrences at each time bin for each neuron.

For each image and animal, the 10 recordings were split into 8 training trials and 2 held-out trials per image. This procedure generates 20 stimulus-specific AEs per animal. The combined test set across images contains 40 trials, and splitting is be repeated with 5 different folds to obtain 5 decoding accuracy value per animal.

Decoding is performed in a generative reconstruction-based manner: for each test spike train, the reconstruction loss is computed for all 20 stimulus-specific AEs. The predicted stimulus corresponds to the AE with the lowest reconstruction loss, implementing a minimum-distance generative decoder. Decoding is performed independently per animal, and accuracy is averaged across animals.

3 Results

3.1 Spiking motif extraction as a function of the number of samples

We first investigated how single-layer AEs trained with different reconstruction losses capture underlying spiking patterns, focusing on the effect of the number of training samples. For these experiments, 100 different spiking motif kernels were generated, but each AE was trained on a single kernel at a time. Each kernel has size $N \times T$, with $N = 260$ neurons (matching the average number recorded per animal) and $T = 250$ time bins. The standard deviation of spike timing jitter was set to 4% of T and additive

noise accounting for 10% of the spiking activity is injected. For each kernel, a variable number of spike train samples was drawn to form the training set for the AE.

Figure 2 summarizes the results as a function of the number of training samples. The left panel shows the similarity value, which considers both spike timing and precision. MSE-trained AEs converge to the true Gaussian distributions when the number of samples is high (green curve), capturing both mean timings and temporal precision of the spikes. By contrast, EMD-trained AEs do not fully reproduce the Gaussian width when the number of samples increases (the blue curve does not converge to 1). However, for a reduced number of samples (below 20), EMD-trained AEs obtain better similarity values.

The middle panel shows mean timing similarity, which considers only the relative spike timing by comparing AE weights to an idealized kernel with Dirac delta peaks at the mean spike times. Here, EMD-trained AEs outperform MSE across all sample sizes, suggesting that the temporal sequence itself is represented more accurately, independent of spike precision.

Finally, the right panel shows the mean absolute timing difference. For each neuron, the peak spike time in the learned motif was estimated by fitting a 1D Gaussian to the weights. Both MSE- and EMD-trained AEs converge to near-zero absolute timing error as the number of samples increases, confirming that both methods accurately recover the mean spike timing when sufficient data are available. For small sample sizes (fewer than 30 trials), EMD provides an advantage, producing more reliable estimates of spike timing compared to MSE.

These results illustrate a key trade-off: MSE excels at capturing the full spike distribution, including temporal precision, but requires large datasets. EMD, by contrast, prioritizes the temporal alignment of spikes, yielding sharper, Dirac-like motifs. The use of this loss shows a clear advantage over MSE when training data are limited, while still converging to the correct mean spike timing in large datasets.

3.2 Robustness of the different methods

We next assessed the robustness of the two AE training approaches to different types of noise in the spike trains. For these simulations, both the training and testing sets contained 30 samples, and each procedure was repeated 100 times with newly generated datasets.

Figure 3 illustrates performance under four types of noise introduced via the generative model (Section 2.3): (A) temporal jitter, (B) neuron participation dropout, (C) time warping of sequences, and (D) additive background spikes.

Performance was evaluated using the mean absolute timing difference between the learned motif and the ground-truth sequence. This measure focuses on temporal alignment and is independent of spike precision, providing a fair comparison of the two losses (see Figure 2-(right)).

Figure 3-(A) shows that EMD-trained AEs are more robust to temporal jitter: as jitter increases, MSE-trained AEs degrade faster, while EMD maintains lower absolute timing errors. Similarly, for time warping (Figure 3-(B)), EMD consistently outperforms MSE, although variability across simulations increases with higher warping.

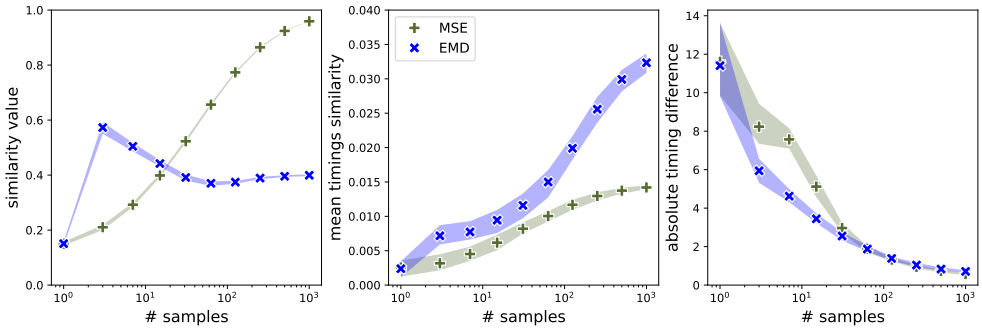


Fig. 2 Performance of single-layer AEs in spiking pattern extraction as a function of the number of training samples. Left: Cosine similarity between AE weights and the ground-truth described by Gaussian distributions, capturing both spike timing and temporal precision. MSE-trained AEs (green) converge to the true distributions for large sample sizes, while EMD-trained AEs (blue) do not fully capture the motifs with their precisions. For small datasets, EMD achieves higher similarity values. **Middle:** Mean timing similarity, which compares AE weights to kernels with Dirac delta peaks at mean spike times, focusing only on temporal alignment. EMD-trained AEs outperform MSE across all sample sizes, indicating sharper representations around mean spike timings. **Right:** Mean absolute timing difference, computed by fitting a 1D Gaussian to each neuron's learned weights and evaluating its absolute temporal distance from the true spike timing. Both methods converge to near-zero error as sample size increases, but EMD provides more reliable spike timing estimates for fewer than 30 trials.

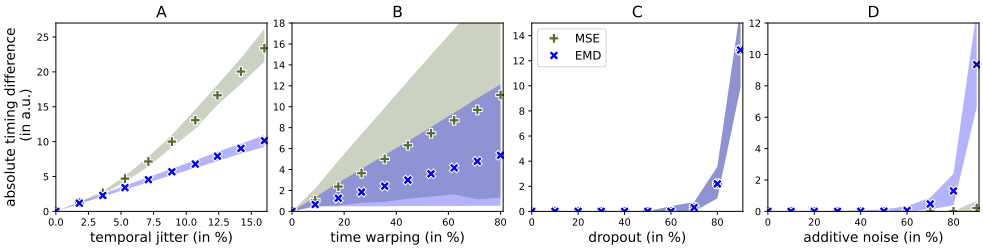


Fig. 3 Robustness of spiking pattern extraction to different noise types. Performance is measured using the mean absolute timing difference between the learned motif and the ground-truth sequence. Four types of noise are considered: (A) temporal jitter, (B) time warping of sequences, (C) neuron participation dropout, and (D) additive background spikes. EMD-trained AEs are more robust to temporal jitter and sequence warping, while MSE-trained AEs handle additive noise more effectively. Both methods show similar robustness to neuron dropout. Error bars indicate the 5^{th} and 95^{th} percentiles across 100 simulation repetitions.

For neuron dropout (Figure 3-(C)), both methods exhibit similar robustness, with errors increasing notably only above $\approx 60\%$ dropout.

Finally, additive background spikes (Figure 3-(D)) affect EMD-trained AEs more than MSE-trained ones, which remain stable even at high levels of unstructured activity. EMD performance starts to degrade noticeably beyond $\approx 60\%$ additive noise.

Overall, these results highlight complementary strengths: EMD excels at preserving temporal structure under jitter and sequence warping, while MSE provides resilience to

additive Poisson noise, illustrating the trade-offs involved in choosing a reconstruction loss for spike motif extraction.

3.3 Decoding performance on Allen Institute Visual Coding data

We next evaluated the ability of stimulus-specific AEs to support decoding of visual inputs from neural activity recorded in the Allen Institute dataset (Section 2.5). Decoding was performed by comparing each test trial to the reconstruction of all 20 stimulus-specific AEs: the stimulus associated with the AE producing the lowest reconstruction loss was selected as the predicted label. Importantly, the reconstruction loss used for decoding could differ from the loss used to train the AE.

Figure 4 summarizes decoding performance across the four possible combinations of training and decoding losses.

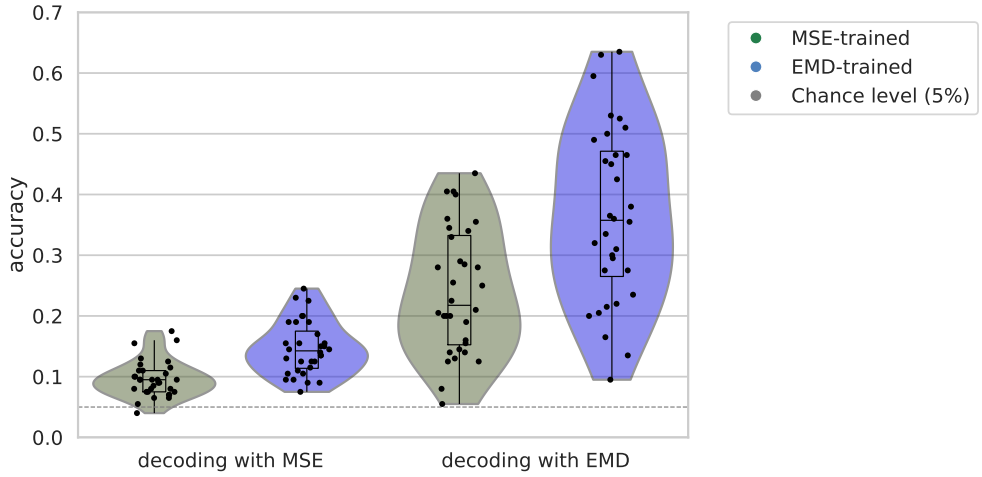


Fig. 4 Decoding accuracy for the Allen Institute Visual Coding dataset. Decoding accuracy values for the 4 different methods applied on the 32 different animals of the dataset. On the left, decoding performance of the MSE-based decoder are above chance level with better performance when AEs are trained with the EMD (blue) than with MSE (green). When EMD is used for decoding (right distributions), accuracy is increased for both training losses. Best decoding performance is obtained when EMD is used both for training and decoding.

All methods obtain, above than chance results, and spike trains are normalized to remove information about firing rates. These results highlight that decoding can be performed based on the temporal structure of the spike trains, beyond firing rates. In Figure B1, spikes of the testing sets are shuffled in time to validate that decoding relies effectively on the temporal struture of spike trains. Indeed, decoding accuracy drops at chance level for all methods when spikes are shuffled along the temporal axis.

In addition, using EMD as a decoding metric consistently improves performance compared to MSE, regardless of the training loss, emphasizing that EMD is better

suited to decode spikes trains thanks to stereotypical spiking motifs learned by the different AEs.

For both decoding losses, training with EMD improves accuracy as well showing that learning spiking motifs benefits from the use of this loss.

Overall, combining EMD-based training and EMD-based decoding maximizes decoding accuracy, yielding more than three times higher performance than the conventional MSE-based approach.

4 Discussion

In this study, we evaluated the use of the EMD as a loss function for extracting spiking motifs using a single-layer AE, and compared its performance to the MSE on both synthetic data and neural recordings from the Allen Institute. Our results show that EMD-based training is more robust to temporal jitter and sequence warping, whereas MSE-based training is less sensitive to additive noise. While MSE-trained AEs perform well with very large datasets, this scenario is often not applicable in typical experimental conditions. With limited samples, EMD-trained AEs produce more reliable estimates of aligned spiking motifs. Applied to mouse visual cortex recordings, EMD-trained AEs reveal temporal sequences that are more discriminative for visual stimuli than those learned using MSE.

4.1 Optimal transport theory as a tool for extracting spiking motifs

Previous work has demonstrated that SpikeShip, a dissimilarity metric based on the EMD, provides better clustering of spike trains than comparisons based on average firing rates [10]. SpikeShip is agnostic to firing rates, and the resulting clusters correspond closely to different visual stimuli. This suggests that spatiotemporal motifs carry meaningful information beyond firing rates and that EMD is well suited to capture it.

Here, we extend this principle to motif extraction using a single-layer AE. In synthetic datasets, EMD-based training reliably recovers the structured temporal sequences used to generate spike trains. The EMD computes the distance between normalized spike trains as probability distributions, comparing cumulative distributions for each neuron and averaging across neurons. Intuitively, this captures how far a spike is shifted relative to its expected timing: the error grows continuously with increasing temporal offset (Figure 1-(B)).

Critically, the EMD enables cross-bin comparisons, interpolating between spikes rather than assigning weight solely to discrete time bins. For example, if spikes occur at two distinct times, t_1 and t_2, the EMD-trained AE will distribute its weights between these times to minimize the temporal distance to both spikes. In contrast, an MSE-trained AE will only place weights on the exact bins corresponding to t_1 and t_2. An MSE-trained AE will have a higher propensity to overfit by learning all the spikes at specific bins in the training set. This can be qualitatively observed in Figure D3 with sharper motifs that are not necessarily shared for training on different datasets linked to the same image. This observation is a specific example of the training with one animal and one specific image. Observing the correlation matrix in Figure D4, reveals

that the variability in the correlation coefficients for AEs trained with the same image is much greater when MSE is used for training. This property makes EMD particularly advantageous when the number of available recordings is small (fewer than 30 trials, as shown in Figure 2) and when spike timings exhibit temporal jitter.

In the Allen Institute dataset, each natural image was presented only 10 times per animal. Despite this limited sample size, EMD-trained AEs successfully extract spiking motifs specific to individual stimuli, outperforming MSE-trained AEs in capturing temporally aligned sequences.

Overall, when neural recordings are limited or spike sequences are affected by temporal jitter or stretching, the EMD provides a robust metric. By preserving the global structure of spike sequences, it tolerates small temporal shifts, making it more resilient to jitter and warping. These types of variability have been demonstrated in natural movie stimuli [15] and in slowed-down hippocampal replay sequences [16], highlighting its broad applicability for extracting meaningful spiking motifs from complex neural activity.

4.2 Comparison with Other Methods to Analyze Spike Trains

Most existing approaches reveal stimulus-induced spiking motifs by analyzing the time to first spike after stimulus onset [3–7, 17]. This latency-based code is effective for identifying temporal sequences, and prior studies have shown that stimuli are often better encoded by the timing of the first spike than by firing rates. However, focusing only on the first spike can miss more complex motifs involving multiple spikes per neuron with precise timing.

The AE developed here is more flexible: it can represent motifs with multiple spikes per neuron, without restricting analysis to a single spike. Similar considerations apply to frequent itemset mining algorithms that detect repeating motifs within fixed temporal bins [8, 9, 18]. Bin size and spike jitter can severely impact these methods: jitter larger than a bin can lead to misclassification of motifs. In contrast, using the EMD allows for cross-bin comparisons, making motif extraction robust to small temporal shifts.

An AE with temporal convolution and EMD loss could cluster spiking motifs in spontaneous activity more effectively. This is conceptually related to convolutional NMF approaches [19–21], such as seqNMF [20], which decomposes the data into sequences (analogous to AE weights) representing repeating spatiotemporal motifs, and temporal loadings (analogous to latent variables) representing motif occurrences. seqNMF relies on a bin-by-bin Frobenius norm between the reconstruction and the signal, making it susceptible to jitter or overfitting, particularly with small datasets and many parameters. In contrast, the EMD-trained AE uses tied weights and a cross-bin metric, reducing the number of parameters and mitigating overfitting. Additionally, seqNMF optimizes the sequence and the temporal loadings independently. This is equivalent to training the latent variables and the AE weights separately, which makes the method more prone to overfitting by increasing the number of parameters. Extending our approach with temporal convolution and multiple latent variables could provide an efficient, fully unsupervised, algorithm for extracting spiking motifs independent of stimulus onset.

4.3 Temporal Code in the Allen Institute Dataset

Previous work with SpikeShip [10] has shown that temporal sequences carry meaningful stimulus-related information and outperform rate-based clustering. Our study, allows to extract interpretable motifs directly from spike train. Key differences include: (1) SpikeShip is a dissimilarity metric, whereas the AE learns a generative representation of the motifs and (2) SpikeShip uses global temporal alignment, which could be mimicked in AEs via temporal convolution.

Overall, results on this dataset suggest that spatiotemporal motifs contribute stimulus-specific information, and even with a simple AE architecture, meaningful motifs can be extracted that capture temporal structure orthogonal to firing rates.

4.4 Limitations and Future Perspectives

Our simulations demonstrate that EMD-trained AEs can reliably capture the mean timings of spikes within structured spiking motifs. However, unlike MSE, the EMD-based method does not converge to the full ground truth distribution representing both spike timing and temporal precision (Figure 2-(left)). Therefore, if precise temporal precision is required and sufficient recordings are available, MSE or another point-by-point metric may be more appropriate.

A key limitation of our approach arises from normalization along the temporal axis. This step is required to preserve a well-defined EMD formalism. Moreover, spikes from neurons with high average firing rates contribute less after normalization, leading the AE to prioritize neurons with lower firing rates. While normalization is necessary to focus on temporal structure rather than firing rate, it introduces a bias, and the choice of whether to account for firing rate imbalances should be considered carefully.

The AE architecture also imposes constraints. Our study uses a single-layer AE with positive, tied weights, limiting the model to linear feature extraction. Extensions using multi-layer or non-linear AEs could capture more complex motifs, while preserving interpretability by maintaining positivity constraints on the decoding weights and positive latent variables making it equivalent to NMF.

Finally, our study restricts training to a single spiking motif with fixed timing, avoiding temporal convolution. While temporal convolution could extract motifs and their occurrences with time invariance, the EMD's sensitivity to background activity makes it challenging to apply over long recordings with sparse motifs. The cumulative reconstruction error in the absence of structure can be mitigated by computing losses on smaller windows. Future work should also investigate overlapping motifs and implement sliding windows to build a fully unsupervised algorithm for detecting spiking motifs in spontaneous activity without stimulus locking.

Supplementary information. Supplementary information is provided in the Appendix section.

Acknowledgements. Thanks to the NeuroAI team in CerCo and specifically Victor Boutin for useful discussions during the elaboration of this work.

Declarations

Author Contributions. Antoine Grimaldi, Boris Sotomayor-Gómez, Matthieu Gilson, Laurent Perrinet and Martin Vinck contributed to the conceptualization and methodology of the study. Andrea Alamia, Martin Vinck, and Matthieu Gilson carried out the funding acquisition and supervision. Formal analysis and investigation were performed by Antoine Grimaldi, Martin Vinck and Andrea Alamia. Boris Sotomayor-Gómez curated and preprocessed the Visual Coding dataset. All authors contributed to the writing and reviewing of the manuscript and approve the final manuscript.

Funding. A.A. was funded by the European Union under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101075930). The copyright holder for this is of the author(s) only and does not necessarily reflect those of the European Union or the European Research Council (ERC). Neither the European Union nor the granting authority can be held responsible for them.

M.G. received support from the French government under the France 2030 investment plan, under the agreement Chaire de Professeur Junior (ANR-22-CPJ2-0020-01) and as part of the Initiative d'Excellence d'Aix-Marseille Université -A*MIDEX (AMX-22-CPJ-01).

M.V. was supported by an ERC starting grant (850861) SPATEMP; DFG VI Grants (908/5-1 and 908/7-1); an NWO VIDI Grant; and the Dutch Brain Interface Initiative (DBI2).

Code Availability. This work is made reproducible. The code reproducing the manuscript and all figures is available on https://github.com/AntoineGrimaldi/wassa/tree/main/results_paper. It also contains supplementary figures and results.

Data and Materials Availability. Not applicable.

Conflict of interest. Not applicable.

Ethics approval and consent to participate. Not applicable.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Appendix A Activation function of the autoencoder minimized with EMD

It is straightforward to demonstrate that the dot product, or convolution in a more general case, used in AEs as the activation function, minimizes the MSE when weights are tied. The variable X is an $N \times T$ matrix, and the estimation of this input by the AE is given by the following formula:

$$\hat{X} = \sum_k a_k W_k = A \cdot W \tag{A1}$$

where a_k are the activations of the different associated sets of weights W_k used for reconstruction. Then, the MSE is given by:

$$\text{MSE}(X, \hat{X}) = \frac{1}{NT} \sum_{i=1}^{NT} (X - \hat{X})_i^2 = \frac{1}{NT} \sum_{i=1}^{NT} (x_i - \hat{x}_i)^2 \tag{A2}$$

By taking the partial derivative of the MSE with respect to a_k one gets:

$$\begin{aligned} \frac{\partial \text{MSE}}{\partial a_k} &= -\frac{2}{NT} \sum_{i=1}^{NT} (X - \hat{X})_i \cdot \frac{\partial \hat{X}_i}{\partial a_k} \\ &= -\frac{2}{NT} \sum_{i=1}^{NT} (X - \sum_k a_k W_k)_i W_{k,i} \\ &= -\frac{2}{NT} \sum_{i=1}^{NT} (X \cdot W_k - W_k \cdot \sum_{k'} a_{k'} W_{k'})_i \end{aligned}$$

In order to reach a minimum for all the data points, one wants the sum to be 0 for all points n. One solution is :

$$\begin{aligned} X \cdot W_k - W_k \cdot \sum_{k' \neq k} a_{k'} W_{k'} - a_k ||W_k||^2 = 0 \\ \Leftrightarrow \quad a_k = \frac{(X - \sum_{k' \neq k} W_{k'}) \cdot W_k}{||W_k||^2} \end{aligned} \tag{A3}$$

In the case of a single kernel ($k = 1$), the dot product divided by the L^2-norm of the weights is recognizable. This demonstration can easily be extended to a convolution paradigm and shows that the dot product or convolution is optimal for MSE minimisation as an activation function.

There is no guarantee that this activation function is also appropriate to minimize the EMD. We can make the same reasoning with the loss function described in Equation 4 to obtain the appropriate activation function:

$$\begin{aligned}\frac{\partial \mathcal{L}(X,\hat{X})}{\partial a_k} &= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{\partial |F_{X_n}(t) - F_{\hat{X}_n}(t)|}{\partial a_k} \\ &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{F_{X_n}(t) - F_{\hat{X}_n}(t)}{|F_{X_n}(t) - F_{\hat{X}_n}(t)|}\frac{\partial F_{\hat{X}_n}(t)}{\partial a_k}\end{aligned} \tag{A4}$$

Since:

$$F_{\hat{X}_n}(t) = \sum_{t'=1}^{t}\sum_{k} a_k W_{k,n,t'} = \sum_{k} a_k \sum_{t'=1}^{t} W_{k,n,t'} = \sum_{k} a_k F_{W_{k,n}}(t) \tag{A5}$$

Equation A4 simplifies to:

$$\begin{aligned}\frac{\partial \mathcal{L}(X,\hat{X})}{\partial a_k} &= -\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{F_{X_n}(t) - F_{\hat{X}_n}(t)}{|F_{X_n}(t) - F_{\tilde{X}_n}(t)|}F_{W_{k,n}}(t) \\ &= -\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{F_{X_n}(t)F_{W_{k,n}}(t) - F_{W_{k,n}}(t)\sum_{k'} a_{k'}F_{W_{k',n}}(t)}{|F_{X_n}(t) - F_{\hat{X}_n}(t)|}\end{aligned} \tag{A6}$$

With the same reasoning as in Equation A3 and the notation for the cumulative used in Equation 4:

$$\begin{aligned}\sum_{n=1}^{N}\sum_{t=1}^{T} F_{X_n}(t)F_{W_{k,n}}(t) - F_{W_{k,n}}(t)\sum_{k'} a_{k'}F_{W_{k',n}}(t) &= 0 \\ \underrightarrow{F_X}\cdot\underrightarrow{F_{W_k}} - \underrightarrow{F_{W_k}}\cdot\sum_{k'} a_{k'}\underrightarrow{F_{W_{k'}}} &= 0 \\ a_k = \frac{\underrightarrow{F_X}\cdot\underrightarrow{F_{W_k}} - \underrightarrow{F_{W_k}}\cdot\sum_{k'} a_{k'}\underrightarrow{F_{W_{k'}}}}{\left\|\underrightarrow{F_{W_k}}\right\|^2}\end{aligned} \tag{A7}$$

When $k = 1$, this activation corresponds to the dot product of the cumulative of the input and the cumulative of the set of weights along the temporal axis divided by the norm of the cumulative of the weights. It can be easily implemented in our AE to use an activation function more appropriate for this loss. Note that this activation can be extended to the convolution case, which is not the focus of this study.

Appendix B Decoding recordings with spikes times shuffled along the temporal axis

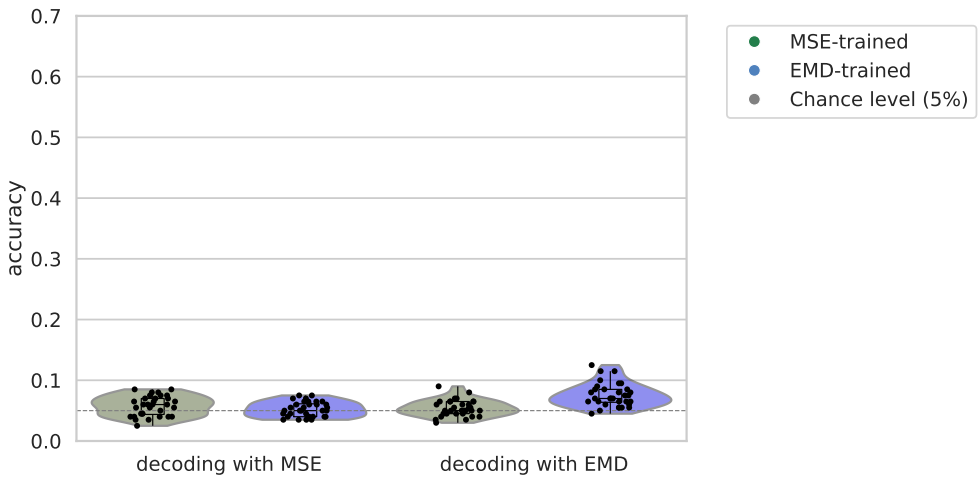


Fig. B1 The decoding accuracy values are shown for the same methods as in Figure 4 when the spikes are shuffled along the temporal axis for the testing set. The same AEs are used for decoding as in Figure 4. The only difference is the random shuffling of spike times, which destroys the temporal information in the testing set. The performance of all methods drops to chance level, showing that the temporal information captured by the weights of the AEs is actually relevant for decoding.

Appendix C Decoding losses

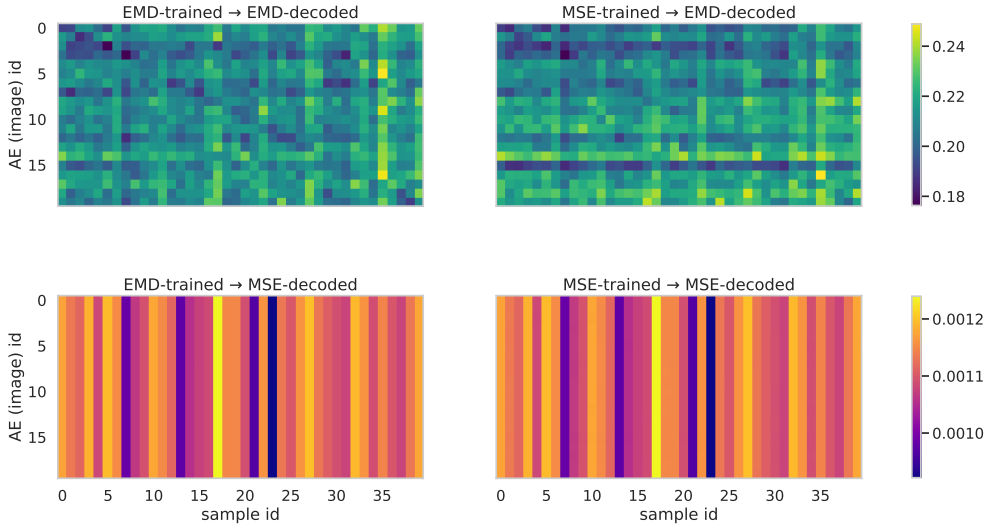


Fig. C2 Reconstruction losses computed on the test trials for the 20 stimulus-specific autoencoders trained on different visual stimuli. Each column corresponds to a test trial and each row to a stimulus-specific AE. Top panels show reconstruction losses computed using the EMD, while bottom panels correspond to losses computed using MSE. With EMD, the reconstruction loss tends to be lowest for the AE associated with the correct stimulus (darker diagonal), supporting accurate decoding. In contrast, when MSE is used for decoding (bottom), vertical stripe patterns appear, indicating that the loss values depend strongly on the specific test trial. This suggests that variability across trials dominates over differences between AEs, likely due to the bin-by-bin nature of the MSE comparison, which is more sensitive to trial-specific fluctuations in spike timing.

Appendix D Spiking motifs

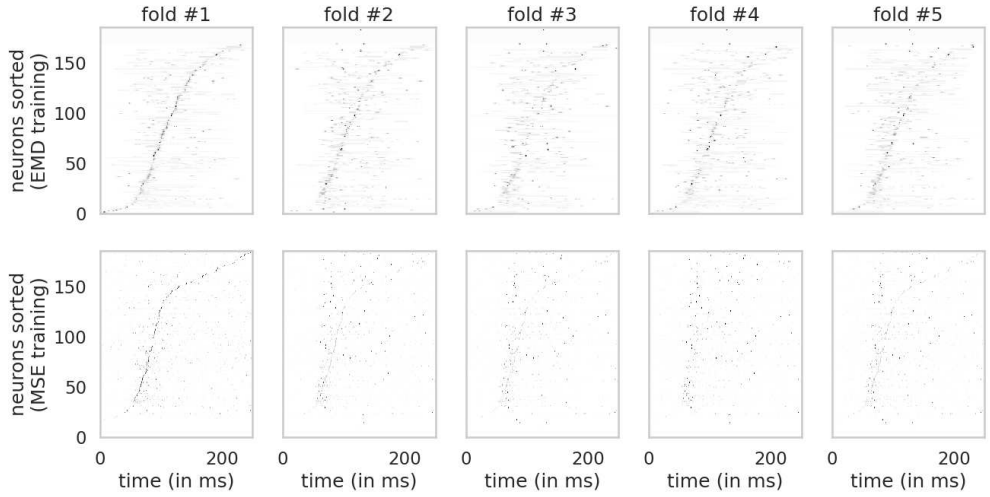


Fig. D3 Example weight patterns for AEs trained on image 19 across different training sets (composed of 8 of the 10 recordings associated with the same image) of recordings on animal 1. Neurons are sorted according to the latency of the maximum weight in the AE trained on training set 1, shown for both EMD (top row) and MSE (bottom row). If a temporal sequence characteristic of a given stimulus is consistently learned, it should be visible across AEs trained on different subsets of data corresponding to the same image. For MSE-trained AEs (bottom row), a precise temporal sequence is visible for training set #1. This structure is not clearly preserved for the other AEs, suggesting that the learned representation may depend strongly on the specific training data, potentially reflecting overfitting to particular trials. In contrast, EMD-trained AEs (top row) capture broader, less sharply defined temporal patterns that remain consistent across training sets when using the same neuron ordering.

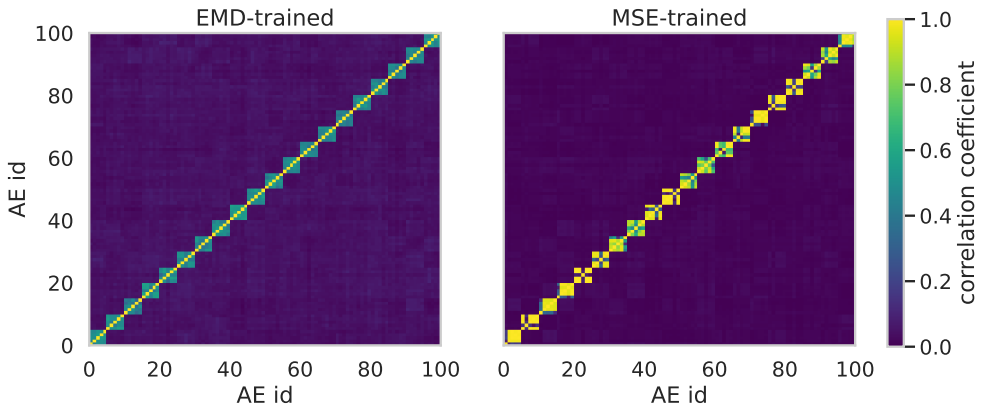


Fig. D4 Example correlation matrix of the trained weights for animal 1. The left panel shows correlations for AEs trained with EMD, and the right panel for AEs trained with MSE. For MSE-trained AEs, correlation coefficients can reach very high values (close to 1) when models are trained on recordings associated with the same image, indicating that the corresponding weight configurations capture highly similar patterns. These high correlations may partly reflect the reuse of similar recordings across training sets for a given image. However, similarity within these clusters can also vary substantially, suggesting that the learned representations are not fully consistent and that the AEs may converge to different solutions depending on the specific training set. In contrast, EMD-trained AEs exhibit lower but more stable within-cluster similarity values.

References

[1] Rudolph, M., Destexhe, A.: Tuning Neocortical Pyramidal Neurons between Integrators and Coincidence Detectors. Journal of Computational Neuroscience **14**(3), 239–251 (2003) https://doi.org/10.1023/A:1023245625896 . Accessed 2024-11-18

[2] Luczak, A., McNaughton, B.L., Harris, K.D.: Packet-based communication in the cortex. Nature Reviews Neuroscience **16**(12), 745–755 (2015) https://doi.org/10.1038/nrn4026 . Publisher: Nature Publishing Group. Accessed 2025-12-10

[3] Johansson, R.S., Birznieks, I.: First spikes in ensembles of human tactile afferents code complex spatial fingertip events. Nature Neuroscience **7**(2), 170–177 (2004) https://doi.org/10.1038/nn1177 . Publisher: Nature Publishing Group. Accessed 2025-12-09

[4] Gollisch, T., Meister, M.: Rapid Neural Coding in the Retina with Relative Spike Latencies. Science **319**(5866), 1108–1111 (2008) https://doi.org/10.1126/science.1149639 . Publisher: American Association for the Advancement of Science. Accessed 2024-02-06

[5] Portelli, G., Barrett, J.M., Hilgen, G., Masquelier, T., Maccione, A., Marco, S.D., Berdondini, L., Kornprobst, P., Sernagor, E.: Rank Order Coding: a Retinal Information Decoding Strategy Revealed by Large-Scale Multielectrode Array Retinal Recordings. eNeuro **3**(3) (2016) https://doi.org/10.1523/ENEURO.0134-15.2016

. Publisher: Society for Neuroscience Section: New Research. Accessed 2025-12-09

[6] Luczak, A., Barthó, P., Harris, K.D.: Spontaneous Events Outline the Realm of Possible Sensory Responses in Neocortical Populations. Neuron **62**(3), 413–425 (2009) https://doi.org/10.1016/j.neuron.2009.03.014 . Publisher: Elsevier. Accessed 2025-12-09

[7] Resulaj, A., Ruediger, S., Olsen, S.R., Scanziani, M.: First spikes in visual cortex enable perceptual discrimination. eLife **7**, 34044 (2018) https://doi.org/10.7554/eLife.34044 . Publisher: eLife Sciences Publications, Ltd. Accessed 2025-12-09

[8] Quaglio, P., Rostami, V., Torre, E., Grün, S.: Methods for identification of spike patterns in massively parallel spike trains. Biological Cybernetics **112**(1-2), 57–80 (2018) https://doi.org/10.1007/s00422-018-0755-0 . Accessed 2024-03-10

[9] Grün, S., Diesmann, M., Grammont, F., Riehle, A., Aertsen, A.: Detecting unitary events without discretization of time. Journal of Neuroscience Methods **94**(1), 67–79 (1999) https://doi.org/10.1016/S0165-0270(99)00126-0 . Accessed 2025-12-09

[10] Sotomayor-Gómez, B., Battaglia, F.P., Vinck, M.: SpikeShip: A method for fast, unsupervised discovery of high-dimensional neural spiking patterns. PLOS Computational Biology **19**(7), 1011335 (2023) https://doi.org/10.1371/journal.pcbi.1011335 . Publisher: Public Library of Science. Accessed 2024-03-10

[11] Vries, S.E., Siegle, J.H., Koch, C.: Sharing neurophysiology data from the Allen Brain Observatory. eLife **12**, 85550 (2023) https://doi.org/10.7554/eLife.85550 . Publisher: eLife Sciences Publications, Ltd. Accessed 2025-12-10

[12] Cornford, J., Pogodin, R., Ghosh, A., Sheng, K., Bicknell, B.A., Codol, O., Clark, B.A., Lajoie, G., Richards, B.A.: Brain-like learning with exponentiated gradients. bioRxiv. Pages: 2024.10.25.620272 Section: New Results (2024). https://doi.org/10.1101/2024.10.25.620272 . https://www.biorxiv.org/content/10.1101/2024.10.25.620272v1 Accessed 2025-10-27

[13] Khatib, A.E., Huang, S., Ghodsi, A., Karray, F.: Nonnegative Matrix Factorization Using Autoencoders And Exponentiated Gradient Descent. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2018). https://doi.org/10.1109/IJCNN.2018.8489242 . ISSN: 2161-4407. https://ieeexplore.ieee.org/abstract/document/8489242 Accessed 2024-12-19

[14] Peyré, G., Cuturi, M.: Computational Optimal Transport. arXiv. arXiv:1803.00567 [stat] (2020). http://arxiv.org/abs/1803.00567 Accessed 2024-03-19

[15] Sotomayor-Gómez, B., Battaglia, F.P., Vinck, M.: Firing rates in visual cortex show representational drift, while temporal spike sequences remain stable. Cell Reports **44**(4) (2025) https://doi.org/10.1016/j.celrep.2025.115547 . Publisher:

Elsevier. Accessed 2025-11-17

[16] Berners-Lee, A., Feng, T., Silva, D., Wu, X., Ambrose, E.R., Pfeiffer, B.E., Foster, D.J.: Hippocampal replays appear after a single experience and incorporate greater detail with more experience. Neuron **110**(11), 1829–1842 (2022) https://doi.org/10.1016/j.neuron.2022.03.010 . Publisher: Elsevier. Accessed 2025-11-25

[17] Xie, W., Wittig, J.H., Chapeton, J.I., El-Kalliny, M., Jackson, S.N., Inati, S.K., Zaghloul, K.A.: Neuronal sequences in population bursts encode information in human cortex. Nature **635**(8040), 935–942 (2024) https://doi.org/10.1038/s41586-024-08075-8 . Publisher: Nature Publishing Group. Accessed 2025-12-09

[18] Stella, A., Quaglio, P., Torre, E., Grün, S.: 3d-SPADE: Significance evaluation of spatio-temporal patterns of various temporal extents. Biosystems **185**, 104022 (2019) https://doi.org/10.1016/j.biosystems.2019.104022 . Accessed 2024-03-10

[19] Peter, S., Kirschbaum, E., Both, M., Campbell, L., Harvey, B., Heins, C., Durstewitz, D., Diego, F., Hamprecht, F.A.: Sparse convolutional coding for neuronal assembly detection. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). ... Accessed 2024-03-11

[20] Mackevicius, E.L., Bahle, A.H., Williams, A.H., Gu, S., Denisenko, N.I., Goldman, M.S., Fee, M.S.: Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience. eLife **8**, 38471 (2019) https://doi.org/10.7554/eLife.38471 . Publisher: eLife Sciences Publications, Ltd. Accessed 2024-10-09

[21] Koshkin, R., Fukai, T.: convSeq: Fast and Scalable Method for Detecting Patterns in Spike Data. arXiv. arXiv:2402.01130 [eess] (2024). http://arxiv.org/abs/2402.01130 Accessed 2024-10-09