Embedding optimization reveals long-lasting history dependence in neural spiking activity

Lucas Rudelt[1*], Daniel González Marx[1], Michael Wibral[2], Viola Priesemann[1,3*]

1 Max-Planck-Institute for Dynamics and Self-Organization, Göttingen, Germany
2 Campus Institute for Dynamics of Biological Networks, University of Göttingen, Göttingen, Germany
3 Bernstein Center for Computational Neuroscience (BCCN) Göttingen

*Corresponding authors: lucas.rudelt@ds.mpg.de, viola.priesemann@ds.mpg.de

Abstract

Information processing can leave distinct footprints on the statistical history dependence in single neuron spiking. Statistical history dependence can be quantified using information theory, but its estimation from experimental recordings is only possible for a reduced representation of past spiking, a so called past embedding. Here, we present a novel embedding-optimization approach that optimizes temporal binning of past spiking to capture most history dependence, while a reliable estimation is ensured by regularization. The approach does not only quantify non-linear and higher-order dependencies, but also provides an estimate of the temporal depth that history dependence reaches into the past. We benchmarked the approach on simulated spike recordings of a leaky integrate-and-fire neuron with long lasting spike-frequency-adaptation, where it accurately estimated history dependence over hundreds of milliseconds. In a diversity of extra-cellular spike recordings, including highly parallel recordings using a Neuropixel probe, we found some neurons with surprisingly strong history dependence, which could last up to seconds. Both aspects, the magnitude and the temporal depth of history dependence, showed interesting differences between recorded systems, which points at systematic differences in information processing between these systems. We provide practical guidelines in this paper and a toolbox for Python3 at `https://github.com/Priesemann-Group/hdestimator` for readers interested in applying the method to their data.

Author summary

Even with exciting advances in recording techniques of neural spiking activity, experiments only provide a comparably short glimpse into the activity of only a tiny subset of all neurons. How can we learn from these experiments about the organization of information processing in the brain? To that end, we exploit that different properties of information processing leave distinct footprints on the firing statistics of individual spiking neurons. In our work, we focus on a particular statistical footprint: How much does a single neuron's spiking depend on its own preceding activity, which we call history dependence. By quantifying history dependence in neural spike recordings, one can, in turn, infer some of the properties of information processing. Because recording lengths are limited in practice, a direct estimation of history dependence from experiments is challenging. The embedding optimization approach that we present in this paper aims at extracting a maximum of history dependence within the limits set by a reliable estimation. The approach is highly adaptive and thereby enables a meaningful comparison of history dependence between neurons with vastly different spiking statistics, which we exemplify on a diversity of spike recordings. In conjunction with recent, highly parallel spike recording techniques, the approach could yield valuable insights on how hierarchical processing is organized in the brain.

Introduction

How is information processing organized in the brain, and what are the principles that govern neural coding? Fortunately, footprints of different information processing and neural coding strategies can be found in the firing statistics of individual neurons, and in particular in the history dependence, the statistical dependence of a single neuron's spiking on its preceding activity.

In classical, noise-less efficient coding, history dependence should be low to minimize redundancy and optimize efficiency of neural information transmission [1, 2, 3]. In contrast, in the presence of noise, history dependence and thus redundancy could be higher to increase the signal-to-noise ratio for a robust code [4]. Moreover, history dependence can be harnessed for active information storage, i.e. maintaining past input information to combine it with present input for temporal processing [5, 6] and associative learning [7]. In addition to its magnitude, the temporal depth of history dependence provides an important footprint of processing at different processing stages in the brain [8]. This is because higher-level processing requires integrating information on longer timescales than lower-level processing [9]. Therefore, history dependence in neural spiking should reach further into the past for neurons involved in higher level processing [8, 10]. Quantifying history dependence could probe these different footprints and thus yield valuable insights on how neural coding and information processing is organized in the brain.

To quantify history dependence in single neuron spiking, one has to compute the mutual information between the spiking of a neuron and its own past [5, 6, 11]. Estimating this mutual information directly from spike recordings, however, is notoriously difficult. The reason is that statistical dependencies may reside in precise spike times, extend far into the past and contain higher-order dependencies. This makes it hard to find a parametric model, e.g. from the family of generalized linear models [12, 13], that is flexible enough to account for the variety of spiking statistics encountered in experiments. Therefore, one typically infers mutual information directly from observed spike trains [14, 15, 16, 17, 18]. The downside is that this requires a lot of data, otherwise estimates can be severely biased [19, 20]. A lot of work has been devoted to finding less biased estimates, either by correcting bias [20, 21, 22, 23], or by using Bayesian inference [24, 25, 26]. Although these estimators alleviate to some extent the problem of bias, a reliable estimation is only possible for a much reduced representation of past spiking, also called past embedding [27]. For example, many studies infer history dependence and transfer entropy by embedding the past spiking using a single bin [18, 28].

While previously most attention was devoted to a robust estimation given a (potentially limited) embedding, we argue that a careful embedding of past activity is crucial. In particular, a past embedding should be well adapted to the spiking statistics of a neuron, but also be low dimensional enough such that reliable estimation is possible. To that end, we here devise an embedding optimization scheme that selects the embedding that maximizes the estimated history dependence, while reliable estimation is ensured by two independent regularization methods.

In this paper, we first provide a short summary of the method employed to estimate history dependence in neural spike trains. A glossary of all the abbreviations and symbols used in this paper can be found at the beginning of the Materials and methods section. In the Results, we test the embedding optimization approach on a tractable benchmark model, and also compare it to existing estimation methods on a variety of experimental spike recordings. Finally, we demonstrate that the approach reveals interesting differences between neural systems, both in terms of the total history dependence, as well as its temporal depth. For the reader interested in applying the method, we provide practical guidelines in the end of the Materials and methods section. The method is readily applicable to highly parallel spike recordings, and a toolbox for Python3 is available online [29].

Methods summary

Definition of history dependence. First, we define the history dependence R in the spiking of a single neuron. We quantify history dependence R based on the mutual information $I(\text{spiking}; \text{past})$

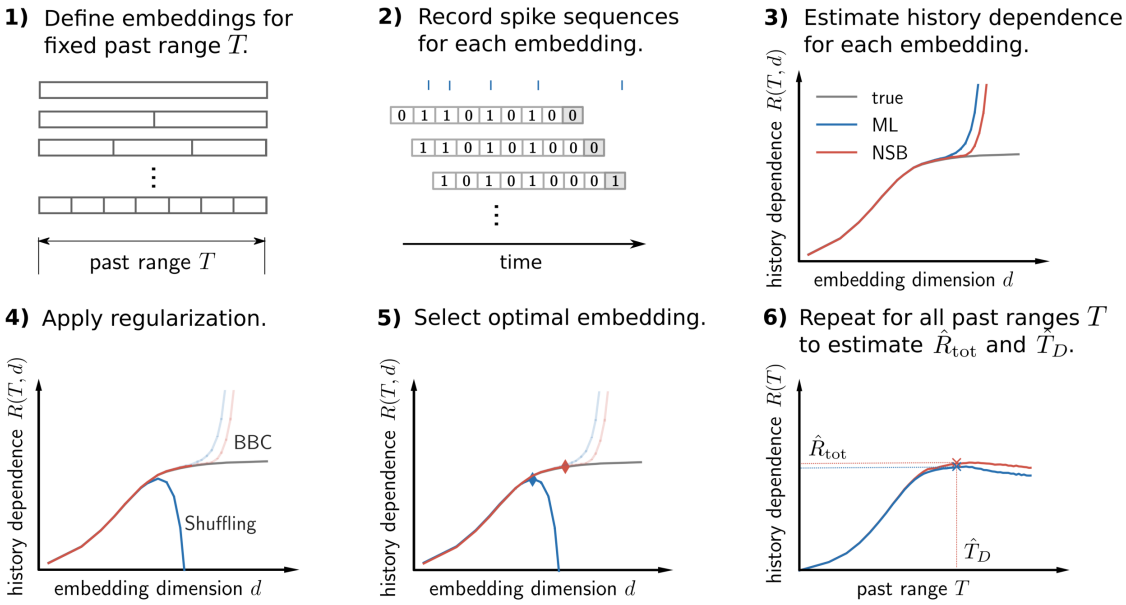

Fig 1. Workflow of past-embedding optimization to estimate history dependence and its temporal depth.

between the current spiking and its own past,

$$R \equiv \frac{I(\text{spiking}; \text{past})}{H(\text{spiking})} = 1 - \frac{H(\text{spiking}|\text{past})}{H(\text{spiking})} \in [0, 1],$$

and normalize it with the Shannon entropy of current spiking $H(\text{spiking})$. Current spiking refers to the firing of a spike in a small time bin $\Delta t = 5\,\text{ms}$, which discretizes the spiking activity in time. Thus the Shannon entropy [30] for current spiking reads

$$H(\text{spiking}) = -p(\text{spike}) \log_2 p(\text{spike}) - (1 - p(\text{spike})) \log_2 (1 - p(\text{spike})),$$

where $p(\text{spike}) = r\Delta t$ is the probability to spike within the time bin Δt for a neuron with average firing rate r. The Shannon entropy $H(\text{spiking})$ quantifies the average information that a spiking neuron could transmit within one bin, assuming no statistical dependencies on its own past. In contrast, the conditional entropy $H(\text{spiking}|\text{past})$ (see Materials and methods) quantifies the average spiking information (in the sense of entropy) that would be transmitted when history dependence on past spiking is taken into account. Note that history dependence can only reduce the average spiking information, i.e. $H(\text{spiking}|\text{past}) \leq H(\text{spiking})$.

The history dependence R accounts for all linear and non-linear as well as higher-order statistical dependencies between current spiking and its own past. To quantify history dependence R, we chose the normalized mutual information, because it can easily be compared across recordings of neurons with very different firing rates. Moreover, it interpolates between the following intuitive extreme cases: $R = 0$ corresponds to independent and $R = 1$ to entirely predictable spiking.

In the following, we summarize the past-embedding approach to estimate history dependence for neural data. The workflow of the approach is illustrated in Fig 1.

Discrete past embedding of spiking activity. In practice, estimating history dependence R from spike recordings is extremely challenging. In fact, if data is limited, a reliable estimation of history dependence is only possible for a reduced representation of past spiking, also called past embedding [27]. Here, we outline how we embed past spiking activity to estimate history dependence from neural spike recordings.

First, we choose a past range T, which defines the time span of the past embedding. For each point in time t, we partition the immediate past window $[t-T, t)$ into d bins and count the number of spikes in each bin. The number of bins d sets the temporal resolution of the embedding. In addition, we let bin sizes scale exponentially with the bin index $j = 1, ..., d$ as $\tau_j = \tau_1 10^{(j-1)\kappa}$ (Fig 2A). A scaling exponent of $\kappa = 0$ translates into equal bin sizes, whereas for $\kappa > 0$ bin sizes increase. For fixed d, this allows to obtain a higher temporal resolution on recent past spikes by decreasing the resolution on distant past spikes.

The past window $[t-T, t)$ of the embedding is slided forward in steps of Δt through the whole recording with recording length T_{rec}, starting at $t = T$. This gives rise to $N = (T_{\text{rec}} - T)/\Delta t$ measurements of current spiking in $[t, t+\Delta t]$, and of the number of spikes in each of the d past bins (Fig 2B). We chose to use only binary sequences of spike counts to estimate history dependence. To that end, a count of 1 was chosen for a spike count larger than the median spike count over the N measurements in the respective past bin. A binary representation drastically reduces the number of possible past sequences for given number of bins d, such that history dependence can be estimated even from short recordings.

Estimation of history dependence for discrete past embeddings. To estimate history dependence R, one has to estimate the probability of a spike occurring together with different past sequences. The probabilities π_i of these different joint events i can be directly inferred from the frequencies n_i with which the events occurred during the recording. Without any additional assumptions, the simplest way to estimate the probabilities is to compute the relative frequencies $\hat{\pi}_i = n_i/N$, where N is the total number of observed joint events. This estimate is the maximum likelihood (ML) estimate of joint probabilities π_i for a multinomial likelihood, and the corresponding estimate of history dependence will also be denoted by ML. This direct estimate of history dependence is known to be strongly biased when data is too limited [20, 22]. The bias is typically positive, because, under limited data, probabilities of *observed* joint events are given too much weight. Therefore, statistical dependencies are overestimated. Even worse, the overestimation becomes more severe the higher the number of possible past sequences K. Since K increases exponentially with the dimension of the past embedding d, i.e. $K = 2^d$ for binary spike sequences, history dependence is severely overestimated for high d (Fig 2C). The potential overestimation makes it hard to choose embeddings that represent past spiking sufficiently well. In the following, we outline how one can optimally choose embeddings if appropriate regularization is applied.

Estimating history dependence with past-embedding optimization. Due to systematic overestimation, high-dimensional past embeddings are prohibitive for a reliable estimation of history dependence from limited data. Yet, high-dimensional past embeddings might be required to capture all history dependence. The reason is that history dependence may reside in precise spike times, but also may extend far into the past.

To illustrate this trade-off, we consider a discrete past embedding of spiking activity in a past range T, where the past spikes are assigned to d equally large bins ($\kappa = 0$). We would like to obtain an estimate $\hat{R}(T)$ of the maximum possible history dependence $R(T)$ for the given past range T, with $R(T) \equiv R(T, d \to \infty)$ (Fig 2C). The number of bins d can go to infinity only in theory, though. In practice, we have estimates $\hat{R}(T, d)$ of the history dependence $R(T, d)$ for finite d. On the one hand, one would like to choose a high number of bins d, such that $R(T, d)$ approximates $R(T)$ well for the given past range T. Too few bins d otherwise reduce the temporal resolution, such that $R(T, d)$ is substantially less than $R(T)$ (Fig 2C). On the other hand, one would like to choose d not too large in order to enable a reliable estimation from limited data. If d is too high, estimates $\hat{R}(T, d)$ strongly overestimate the true history dependence $R(T, d)$ (Fig 2C).

Therefore, if the past embedding is not chosen carefully, history dependence is either overestimated due to strong estimation bias, or underestimated because the chosen past embedding was too simple.

Here, we thus propose the following *past-embedding optimization* approach: For a given past range T, select embedding parameters d^*, κ^* that maximize the estimated history dependence $\hat{R}(T, d, \kappa)$, while overestimation is avoided by an appropriate regularization. This yields an embedding-optimized

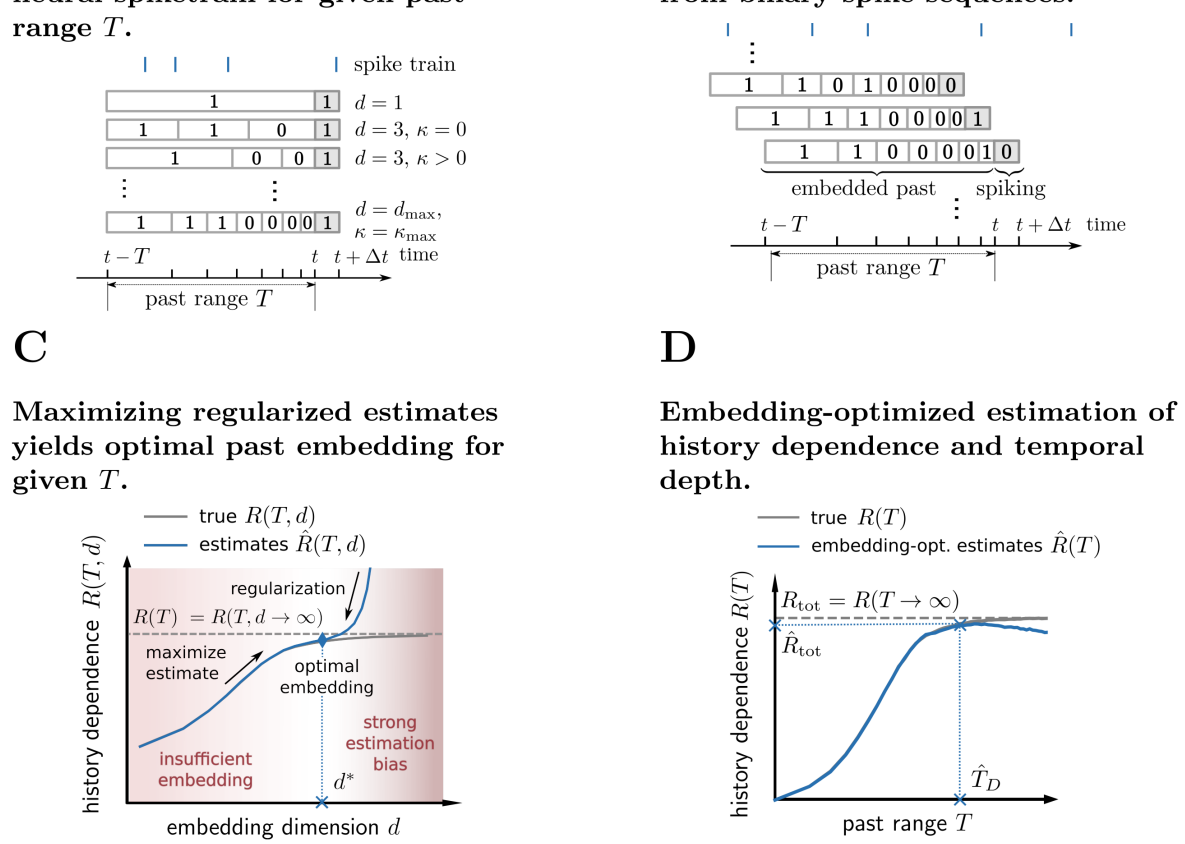

Fig 2. Illustration of embedding optimization to estimate history dependence and its temporal depth. A History dependence R is estimated from the observed joint statistics of current spiking in a small time bin $[t + \Delta t]$ (dark grey) and the embedded past, i.e. a binary sequence representing past spiking in a past window $[t - T, t)$. We systematically vary the number of bins d and bin sizes for fixed past range T. Bin sizes scale exponentially with bin index and a scaling exponent κ to reduce resolution for spikes further into the past. **B** The joint statistics of current and past spiking are obtained by shifting the past range in steps of Δt and counting the resulting binary sequences. **C** Finding a good choice of embedding parameters (e.g. embedding dimension d) is challenging: When d is chosen too small, the true history dependence $R(T)$ (dashed line) is not captured appropriately (insufficient embedding) and underestimated by estimates $\hat{R}(T, d)$ (blue solid line). When d is chosen too high, estimates $\hat{R}(T, d)$ are severely biased and $R(T, d)$, as well as $R(T)$, are overestimated (biased regime). Past-embedding optimization finds the optimal embedding parameter d^* that maximizes the estimated history dependence $\hat{R}(T, d)$ subject to regularization. This yields a best estimate $\hat{R}(T)$ of $R(T)$ (blue diamond). **D** Estimation of history dependence $R(T)$ as a function of past range T. For each past range T, embedding parameters d and κ are optimized to yield an embedding-optimized estimate $\hat{R}(T)$. From estimates $\hat{R}(T)$ we obtain estimates of the temporal depth $\hat{T}_D$, as well as the total history dependence $\hat{R}_{\text{tot}}$. For high past ranges T estimates $\hat{R}(T)$ decrease, because reliable estimation requires a reduced temporal resolution.

estimate $\hat{R}(T) = \hat{R}(T, d^*, \kappa^*)$ of the true history dependence $R(T)$. In terms of the above example, past-embedding optimization selects the optimal embedding dimension d^*, which provides the best

lower bound $\hat{R}(T) = \hat{R}(T, d^*)$ to $R(T)$ (Fig 2C).

Since we can anyways provide only a lower bound, regularization only has to ensure that estimates $\hat{R}(T, d, \kappa)$ are either unbiased, or a lower bound to the observable history dependence $R(T, d, \kappa)$. For that purpose, in this paper we introduce a Bayesian bias criterion (BBC) that selects only unbiased estimates. In addition, we use an established bias correction, the so called Shuffling estimator [23] that, within leading order of the sample size, is guaranteed to provide a lower bound to the observable history dependence (see Materials and methods for details).

Together with these regularization methods, the embedding optimization approach enables complex embeddings of past activity without the risk of overestimation.

Estimation of temporal depth and total history dependence. In the previous steps, we focused on the estimation of history dependence $R(T)$ for embeddings with a fixed past range T. Here, we describe how we use these estimates to estimate the temporal depth of history dependence, i.e. the time span over which neural spiking depends on its own history, as well as the total history dependence. The temporal depth T_D we defined as the *minimal past range* for which the total history dependence $R_{\text{tot}} \equiv R(T \to \infty)$ is captured. The temporal depth thus quantifies how far history dependence in neural spiking reaches into the past.

Using the embedding-optimized estimates $\hat{R}(T)$, the temporal depth was estimated by the past range $\hat{T}_D$ for which $\hat{R}(T)$ saturated within errorbars (Fig 2D). Errorbars were obtained by bootstrapping, and saturation was determined when an estimate $\hat{R}(T)$ surpassed the overall highest estimate minus the standard deviation $\hat{R}_{\max} - \sigma_{\hat{R}_{\max}}$ (Materials and methods). Taking the standard deviation into account makes estimates of the temporal depth more robust to statistical fluctuations in estimates of the history dependence $\hat{R}(T)$.

Based on the estimated temporal depth $\hat{T}_D$, we estimated the total history dependence $\hat{R}_{\text{tot}}$ by averaging $\hat{R}(T)$ over past ranges $T \in [\hat{T}_D, T_{\max}]$ that are higher than the temporal depth, but also lower than $T_{\max}$. The upper limit at the past range $T_{\max}$ excludes estimates that are systematically underestimated due to limited resolution for high past ranges (Materials and methods).

Results

In the first part, we benchmark the approach using a tractable neuron model. In the second part, we compare it to existing estimation methods on a variety of experimental spike recordings, and arrive at a best practice solution. In the last part, we demonstrate that the approach reveals interesting differences in history dependence between experimental systems.

Embedding optimization can capture long-lasting history dependence for a benchmark spiking neuron model

On a benchmark spiking neuron model, we first demonstrate that without optimization and proper regularization, past embeddings are likely to capture much less history dependence, or lead to estimates that severely overestimate the true history dependence. We then validate that embedding optimization captures the model's history dependence for hundreds of milliseconds, while being robust to systematic overestimation. As a model we chose a generalized leaky integrate-and-fire neuron with spike frequency adaptation, whose parameters were fitted to experimental data [3, 31]. The neuron was driven with a constant input current to achieve an average firing rate of 4 Hz. The model neuron was chosen, because it is equipped with a spike adaptation mechanism that lasts over 20 seconds, and the ground truth of the total history dependence R_{tot} can be directly computed from sufficiently long simulations (Materials and methods). In addition, we showed that the neuron model can be well approximated by a generalized linear model (GLM). By fitting a GLM, we could thus faithfully estimate the true value of history dependence $R(T, d, \kappa)$ for any past embedding T, d, κ (Materials and methods). In the following, estimates are shown for a simulated recording of 90 minutes, whereas GLM estimates were computed on a 900 minute recording.

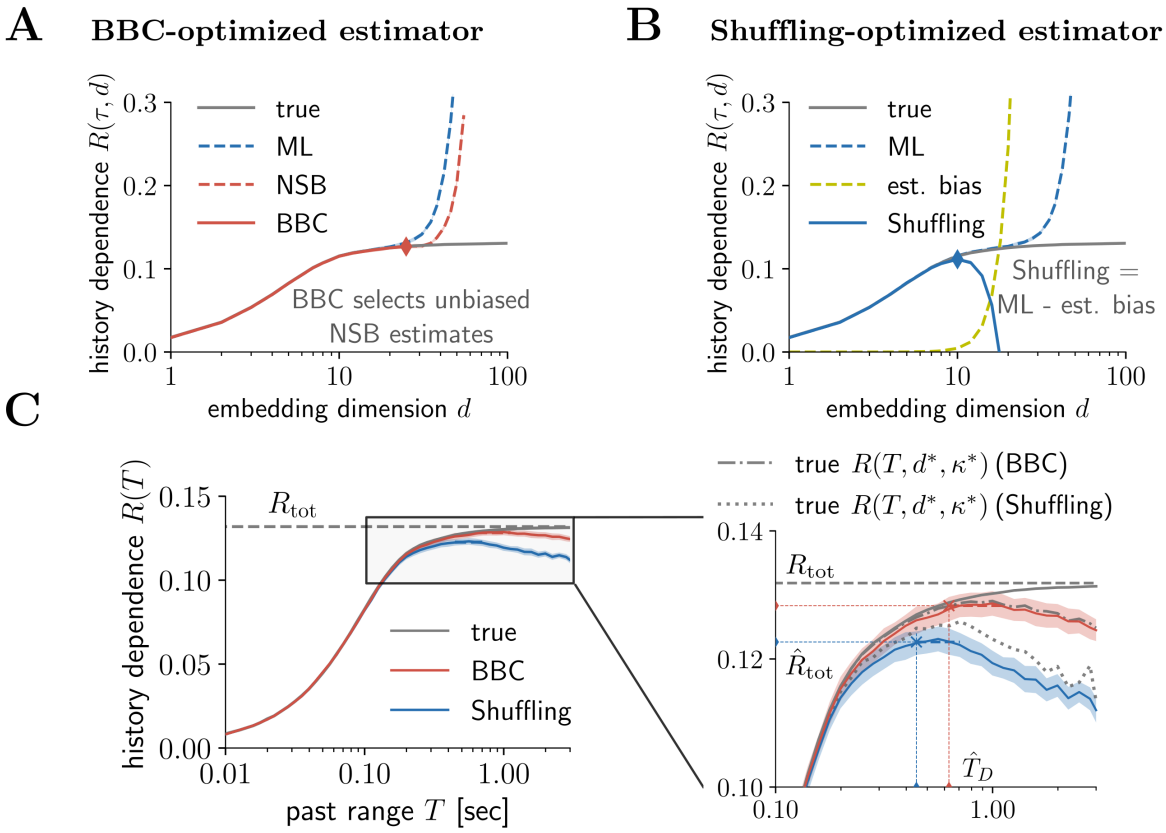

Fig 3. Embedding optimization accurately estimates history dependence for a generalized leaky integrate-and-fire neuron with long-lasting spike frequency adaptation [3, 31]. A For illustration, history dependence $R(\tau, d)$ was estimated on a simulated 90 minute recording for different embedding dimensions d and a fixed bin width $\tau = 20\,\text{ms}$. Maximum likelihood (ML) [20] and Bayesian (NSB) [25] estimators display the insufficient embedding versus estimation bias trade-off: For small embedding dimensions d, the estimated history dependence is much smaller, but agrees well with the true history dependence $R(\tau, d)$ for the given embedding. For larger d, the estimated history dependence $\hat{R}(\tau, d)$ increases, but when d is too high ($d > 20$), it severely overestimates the true $R(\tau, d)$. The Bayesian bias criterion (BBC) selects NSB estimates $\hat{R}(\tau, d)$ for which the difference between ML and NSB estimate is small (red solid line). All selected estimates are unbiased and agree well with the true $R(\tau, d)$ (grey line). Thus, embedding optimization selects the highest, yet unbiased estimate (red diamond). **B** The Shuffling estimator (blue solid line) subtracts estimation bias on surrogate data (yellow dashed line) from the ML estimator (blue dashed line). Since the surrogate bias is higher than the systematic overestimation in the ML estimator (difference between grey and blue dashed lines), the Shuffling estimator is a lower bound to $R(\tau, d)$. Embedding optimization selects the highest estimate, which is still a lower bound (blue diamond). For A and B, shaded areas indicate 2 standard deviations of the estimates obtained from 50 repeated simulations. **C** Embedding optimized estimates $\hat{R}(T)$ yield accurate estimates of the model neuron's history dependence $R(T)$ for hundreds of milliseconds. The zoom-in (right panel) shows robustness of both regularization methods: For all T the model neuron's $R(T)$ lies within errorbars (BBC), or consistently above the Shuffling estimator that provides a lower bound. Here, the model's $R(T)$ was computed for the optimized embedding parameters $d^*, \kappa*$ that were selected via BBC or Shuffling, respectively (dashed lines). Shaded areas indicate 95 % confidence intervals obtained by bootstrapping, and colored dashed lines indicate past ranges over which estimates $\hat{R}(T)$ were averaged to compute $\hat{R}_{\text{tot}}$.

Without regularization, history dependence is severely overestimated for high-dimensional embeddings. For demonstration, we estimated the history dependence $R(\tau, d)$ for varying numbers of bins d and a constant bin size $\tau = 20\,\text{ms}$ (i.e. $\kappa = 0$ and $T = d \cdot \tau$). We compared estimates $\hat{R}(\tau, d)$ obtained by maximum likelihood (ML) estimation [20], or Bayesian estimation using the NSB estimator [25], with the model's true $R(\tau, d)$.

Both estimators accurately estimate $R(\tau, d)$ for up to $d \approx 15$ past bins. As expected, the NSB estimator starts to be biased at higher d than the ML estimator. For embedding dimensions $d > 30$, both estimators severely overestimate $R(\tau, d)$. Note that two standard deviations are plotted as shaded areas, but are too small to be visible. Therefore, any deviation of estimates from the model's true history dependence $R(\tau, d)$ can be attributed to positive estimation bias, i.e. a systematic overestimation of the true history dependence due to limited data.

The aim is now to identify the largest embedding dimension d^* for which the estimate of $R(\tau, d)$ is not yet biased. A biased estimate is expected as soon as the two estimates ML and NSB start to differ significantly from each other (Fig 3A, red diamond), which is formalized by the Bayesian bias criterion (BBC) (Materials and methods). According to the BBC, all NSB estimates $\hat{R}(\tau, d)$ with d lower or equal to d^* are unbiased (solid red line). We find that indeed all BBC estimates agree well with the true $R(\tau, d)$ (grey line), but d^* yields the largest unbiased estimate.

The problem of estimation bias has also been addressed previously by the so-called Shuffling estimator [23]. The Shuffling estimator is based on the ML estimator and applies a bias correction term (Fig 3B). In detail, one approximates the estimation bias using surrogate data, which are obtained by shuffling of the embedded spiking history. The surrogate estimation bias (yellow dashed line) is proven to be larger than the actual estimation bias (difference between grey solid and blue dashed line). Therefore, Shuffling estimates $\hat{R}(\tau, d)$ provide lower bounds to the true history dependence $R(\tau, d)$. As with the BBC, one can safely maximize Shuffling estimates $\hat{R}(\tau, d)$ over d to find the embedding dimension d^* that provides the largest lower bound to the model's total history dependence R_{tot} (Fig 3B, blue diamond).

Thus, using a model neuron, we illustrated that history dependence can be severely overestimated if the embedding is chosen too complex. Only when overestimation is tamed by one of the two regularization methods, BBC or Shuffling, embedding parameters can be safely optimized to yield better estimates of history dependence.

Optimized embeddings capture the model's true history dependence for hundreds of milliseconds. In the previous part, we demonstrated how embedding parameters are optimized for the example of fixed κ and τ. Now, we optimize all embedding parameters for fixed past range T to obtain embedding-optimized estimates $\hat{R}(T)$ of $R(T)$. In particular, we test whether estimates $\hat{R}(T)$ agree well with the model's true history dependence $R(T)$ (see Materials and methods for details on how we obtained $R(T)$).

Embedding-optimized estimates $\hat{R}(T)$ were computed for a range of T using either the Bayesian bias criterion (BBC) or the Shuffling estimator. Notably, for both estimators, estimates $\hat{R}(T)$ agree with the true history dependence for up to several hundred milliseconds (Fig 3C). When comparing the two regularization methods (BBC and Shuffling), the BBC approach captures more history dependence.

For both regularization methods the estimates $\hat{R}(T)$ decrease for high T. This is because little additional history dependence is uncovered, whereas the constraint of an unbiased estimation decreases the temporal resolution. Thus for very high past ranges T, the embedding-optimized estimates are considerably below the true history dependence of the underlying model neuron. The estimated temporal depth $\hat{T}_D \approx 630\,\text{ms}$ for BBC is therefore smaller than the true temporal depth, which, based on the true $R(T)$, is larger than 3 seconds (Fig 3C). The true total history dependence of $R_{\text{tot}} = 13.2\%$ is, however, well estimated with $\hat{R}_{\text{tot}} \approx 12.8\%$ for BBC.

Embedding-optimized estimates do not overestimate history dependence despite maximization over complex embeddings. In the previous part, we investigated how much of the true history dependence for different past ranges T (grey solid line) we miss by embedding the spiking

history. An additional source of error is the estimation of history dependence from limited data. In particular, estimates are prone to overestimate history dependence systematically.

To test explicitly for overestimation, we computed the true history dependence $R(T, d^*, \kappa^*)$ for exactly the same sets of embedding parameters T, d^*, κ^* that were found during embedding optimization with BBC (grey dash-dotted line), and the Shuffling estimator (gray dotted line, Fig 3C, zoom-in). We expect that BBC estimates are unbiased, i.e. the true history dependence should lie within errorbars of the BBC estimates (red shaded area) for a given T. In contrast, Shuffling estimates are a lower bound, i.e. estimates should lie below the true history dependence (given the same T, d^*, κ^*). We find that this is indeed the case for all T. Note that this is a strong result, because it requires that the regularization methods work reliably for every single set of embedding parameters used for optimization—otherwise, parameters that cause overestimation would be selected.

Thus, we can confirm that the embedding-optimized estimates do not systematically overestimate the model neuron's history dependence, and are on average lower bounds to the true history dependence. This is important for the interpretation of the results.

Mild overfitting can occur during embedding optimization on short recordings, but can be overcome with cross-validation. We also tested whether the recording length affects the reliability of embedding-optimized estimates, and found very mild overestimation (1–3%) of history dependence for BBC for recordings as short as 3 minutes (supplementary Fig S1). The overestimation is a consequence of overfitting during embedding optimization: variance in the estimates increases for shorter recordings, such that maximizing over estimates selects embedding parameters that have high history dependence by chance. Therefore, the overestimation can be overcome by cross-validation, i.e. optimizing embedding parameters on one half, and computing estimates on the other half of the data (supplementary Fig S1). In contrast, we found that for the model neuron, Shuffling estimates do not overestimate the true history dependence even for recordings as short as 3 minutes (supplementary Fig S1). This is because the effect of overfitting was small compared to the systematic underestimation of Shuffling estimates. Here, all experimental recordings where we apply BBC are long enough (≈ 90 minutes), such that overfitting was neglected in this paper.

Estimates of temporal depth are sensitive to the recording length. Finally, we also tested the impact of the recording length on the value of the estimated total history dependence $\hat{R}_{\text{tot}}$, as well as the temporal depth $\hat{T}_D$. While on recordings of 3 minutes embedding optimization still estimated $\approx 95\,\%$ of $\hat{R}_{\text{tot}}$ that was estimated for 90 minutes, the estimated $\hat{T}_D$ was only half of the temporal depth that was estimated for 90 minutes (supplementary Fig S2). The temporal depth decreases for shorter recordings, because the variance of estimates increases, such that estimates $\hat{R}(T)$ saturate within errorbars for smaller past ranges T. We therefore advice to compare history dependence, and especially $\hat{T}_D$, for recordings of similar recording length. In the following, we explicitly shorten some recordings such that all recordings have approximately the same recording length.

In conclusion, embedding optimization accurately estimated the model neuron's history dependence for past ranges of several hundred milliseconds. Moreover, for all past ranges, embedding-optimized estimates were robust to systematic overestimation. Embedding optimization is thus a promising approach to quantify history dependence and temporal depth in experimental spike recordings.

Embedding optimization reveals long-lasting history dependence in recordings of spiking neurons

Here, we apply embedding optimization to long spike recordings (≈ 90 minutes) from rat dorsal hippocampus layer CA1 [32, 33], salamander retina [34, 35] and in vitro recordings of rat cortical culture [36]. In particular, we compare embedding optimization to other popular estimation approaches, and demonstrate that an exponential past embedding is necessary to estimate history dependence for long past ranges. We conclude with a practical advice on how to estimate history dependence in highly parallel recordings of spiking neurons.

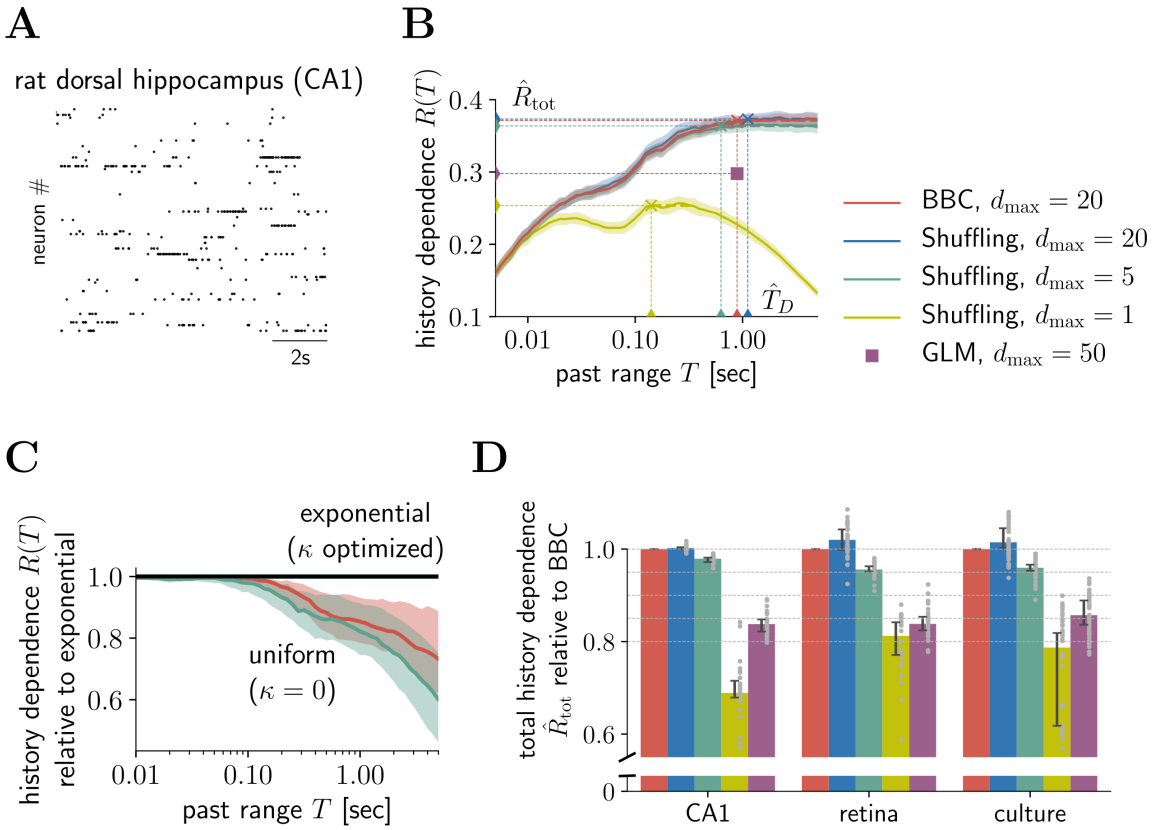

Fig 4. Embedding optimization reveals strong and long-lasting history dependence in experimental spike recordings. A Example of recorded spiking activity in rat dorsal hippocampus layer CA1. **B** Estimated history dependence $\hat{R}(T)$ for various estimators, as well as the estimated total history dependence $\hat{R}_{\text{tot}}$ and temporal depth $\hat{T}_D$ (dashed lines) (example neuron from CA1). Embedding optimization with BBC (red) and Shuffling (blue) for $d_{\max} = 20$ yields consistent estimates. Embedding-optimized Shuffling estimates with a maximum of $d_{\max} = 5$ past bins (green) are very similar to estimates obtained with $d_{\max} = 20$ (blue). In contrast, using a single past bin ($d_{\max} = 1$, yellow), or fitting a GLM for the temporal depth $\hat{T}_D$ found with BBC (violet dot), estimates much lower total history dependence. Shaded areas show standard deviation obtained by bootstrapping, and colored dashed lines indicate past ranges over which estimates $\hat{R}(T)$ were averaged to compute $\hat{R}_{\text{tot}}$. **C** An exponential past embedding is crucial to capture history dependence for high past ranges T. For $T > 100$ ms, uniform embeddings strongly underestimate history dependence. Shown is the median of embedding-optimized estimates $\hat{R}(T)$ with uniform embeddings, relative to estimates obtained by optimizing exponential embeddings, for BBC with $d_{\max} = 20$ (red) and Shuffling with $d_{\max} = 5$ (green). Shaded areas show 95 % percentiles. Median and percentiles were computed over analyzed neurons in CA1 ($n = 28$). **D** Comparison of estimated total history dependence $\hat{R}_{\text{tot}}$ for different estimation and embedding techniques for three different experimental recordings. For each neuron (grey dots), estimates are plotted relative to embedding-optimized estimates for BBC and $d_{\max} = 20$. Embedding optimization with Shuffling and $d_{\max} = 20$ yields consistent but slightly higher estimates than BBC. Strikingly, Shuffling estimates for as little as $d_{\max} = 5$ past bins (green) capture more than 95 % of the estimates for $d_{\max} = 20$ (BBC). In contrast, Shuffling estimates obtained by optimizing a single past bin, or fitting a GLM, are considerably lower. Bars indicate the median and lines indicate 95 % bootstrapping confidence intervals on the median over analyzed neurons (CA1: $n = 28$; retina: $n = 111$; culture: $n = 48$).

Embedding optimization reveals history dependence that is not captured by a generalized linear model or a single past bin. We use embedding optimization to estimate history dependence $\hat{R}(T)$ as a function of the past range T (see Fig 4B for an example neuron from hippocampus layer CA1, and supplementary Figs S6, S7 and S8 for all analyzed neurons). In this example, BBC and Shuffling with a maximum of $d_{\max} = 20$ past bins led to very similar estimates for all T. Notably, embedding optimization with both regularization methods estimated high total history dependence of almost up to 40%, and a temporal depth of almost a second (Fig 4B). This indicates that embedding-optimized estimates capture a substantial part of history dependence also in experimental spike recordings.

Importantly, other common estimation approaches fail to capture the same amount of history dependence (Fig 4B,D). To compare how well the different estimation approaches could capture the total history dependence, we plotted for each neuron the different estimates $\hat{R}_{\text{tot}}$ relative to the BBC estimate (Fig 4D). Embedding optimization with Shuffling yields estimates that agree well with BBC estimates. The Shuffling estimator even yields slightly higher values on the experimental data. Interestingly, embedding optimization with the Shuffling estimator and as little as $d_{\max} = 5$ past bins captures almost the same history dependence as BBC with $d_{\max} = 20$, with a median above 95 % for all recorded systems. In contrast, we find that a single past bin only accounts for 70% to 80% of the total history dependence. A GLM bears little additional advantage with a slightly higher median of $\approx 85\%$. To save computation time, GLM estimates were only computed for the temporal depth $\hat{T}_D$ that was found using BBC (Fig 4B, violet square). The remaining embedding parameters d and κ of the GLM's history kernel were separately optimized using the Bayesian information criterion (Materials and methods). Since embedding and model parameters for the GLM were optimized, we argue that the GLM underestimates history dependence because of its model assumption of no interdependencies between past spikes. Considering that embedding optimization with only five past bins estimates much higher history dependence, we conclude that interdependencies between past events are important, but do not require very high temporal resolution.

Increasing bin sizes exponentially is crucial to estimate history dependence for high past ranges. To demonstrate this, we plotted embedding-optimized BBC estimates $\hat{R}(T)$ using a uniform embedding, i.e. equal bin sizes, relative to estimates obtained with exponential embedding (Fig 4C), both for BBC with $d_{\max} = 20$ (red) and Shuffling with $d_{\max} = 5$ (green). For past ranges $T > 100\,\text{ms}$, estimates using a uniform embedding miss considerable history dependence, which becomes more severe the longer the past range. In the case of $d_{\max} = 5$, a uniform embedding captures around 80 % for $T = 1\,\text{s}$, and only around 60 % for $T = 10\,\text{s}$ (median over analyzed neurons in CA1). Therefore, we argue that an exponential embedding is crucial when assessing the temporal depth of history dependence in neural spiking activity.

Practical advice on how to estimate history dependence. We found that embedding optimization yields an efficient and robust way to estimate history dependence in experimental spike recordings. To leverage the full potential of the approach one should consider an exponential increase of past bin sizes, especially for high past ranges. Interestingly, optimizing embeddings with as few as five past bins is sufficient to capture most history dependence, which strongly reduces computation time and enables embedding optimization for large, highly parallel spike recordings. We therefore give the following practical advice: To estimate history dependence, use the Shuffling estimator and optimize past embeddings with up to five past bins and including a scaling exponent κ (see Practical guidelines in Materials and methods). In the following, we use this approach to estimate the total history dependence and temporal depth for spike recordings from different species and brain areas.

Embedding optimization reveals clear differences in history dependence between recorded systems and individual neurons

Finally, we present results from diverse electrophysiological spike recordings that show interesting differences in history dependence between neurons of different recorded systems. In addition to

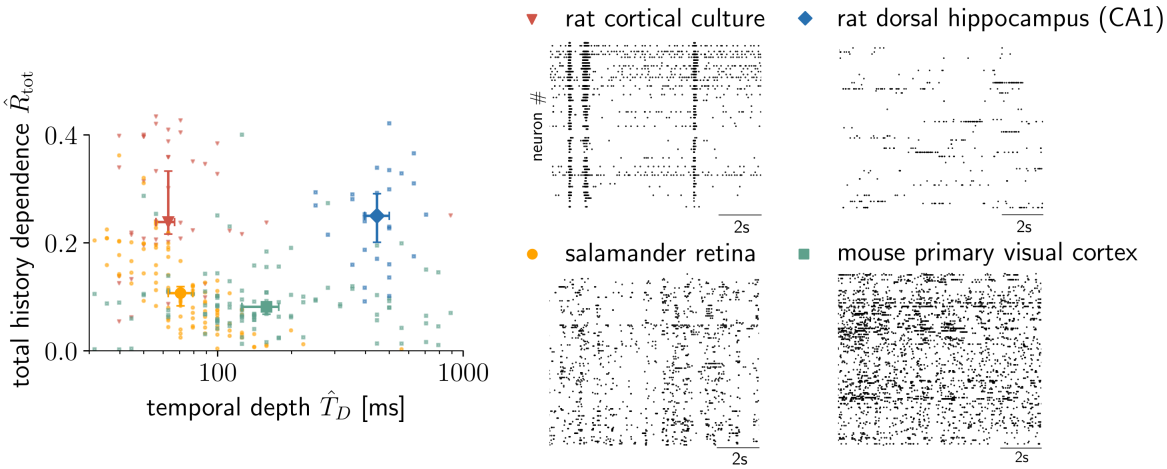

Fig 5. Total history dependence and temporal depth show clear differences between recorded systems. Embedding-optimized Shuffling estimates ($d_{\max} = 5$) of the total history dependence are plotted against the temporal depth for individual neurons (dots) from four different recorded systems (raster plots show spiketrains of the recorded neurons). The analysis shows systematic differences between the recorded systems: Neurons in rat cortical culture and rat dorsal hippocampus layer CA1 have higher median total history dependence than neurons in salamander retina and mouse primary visual cortex. At the same time, neurons in cortical culture and retina show smaller temporal depth than neurons in primary visual cortex, and much smaller temporal depth than neurons in hippocampus layer CA1. Overall, recorded systems are clearly distinguishable when jointly considering the total history dependence and temporal depth. Error bars indicate median over neurons and 95 % bootstrapping confidence intervals on the median.

recordings from rat dorsal hippocampus layer CA1, salamander retina and rat cortical culture, we analyzed neural spike trains recorded in mouse primary visual cortex using the novel Neuropixel probe [37]. Recordings from primary visual cortex were approximately 40 minutes long. Thus, to make results comparable, we analyzed only the first 40 minutes of all recordings.

We find clear differences between the recorded systems, both in terms of the total history dependence, as well as the temporal depth (Fig 5). Neurons in cortical culture and hippocampus layer CA1 have high total history dependence R_{tot} with median over neurons of $\approx 24\,\%$ and $\approx 25\,\%$, whereas neurons in retina and primary visual cortex have typically low R_{tot} of $\approx 11\,\%$ and $\approx 8\,\%$. In terms of temporal depth, neurons in hippocampus layer CA1 display much higher temporal depth T_D with a median of ≈ 450 ms than neurons in cortical culture with median temporal depth of ≈ 60 ms. Similarly, neurons in primary visual cortex have higher T_D with median of ≈ 160 ms than neurons in retina with median of ≈ 70 ms. These differences could reflect differences between early visual processing (retina, primary visual cortex) and high level processing and memory formation in hippocampus, or likewise, between neural networks that are mainly input driven (retina) or exclusively driven by recurrent input (culture). Notably, studying history dependence or the temporal depth of history dependence in isolation would miss differences between recorded systems, whereas considering them jointly allows to distinguish the different systems in terms of history dependence.

In addition to differences between recorded systems, we also find strong heterogeneity of history dependence *within* a single recorded system. Here, we demonstrate this for three different neurons in primary visual cortex (Fig 6, see supplementary Fig S9 for all analyzed neurons in primary visual cortex). In particular, neurons display different signatures of history dependence $R(T)$ as a function of the past range T. For some neurons, history dependence builds up on short past ranges T (e.g. Fig 6A), for some it only shows for higher T(e.g. Fig 6B), and for some it already saturates for very short T (e.g. Fig 6C). We also observed that history dependence can build up on all timescales up to seconds, and that it shows characteristic increases at particular past ranges, e.g. $T \approx 100$ ms and

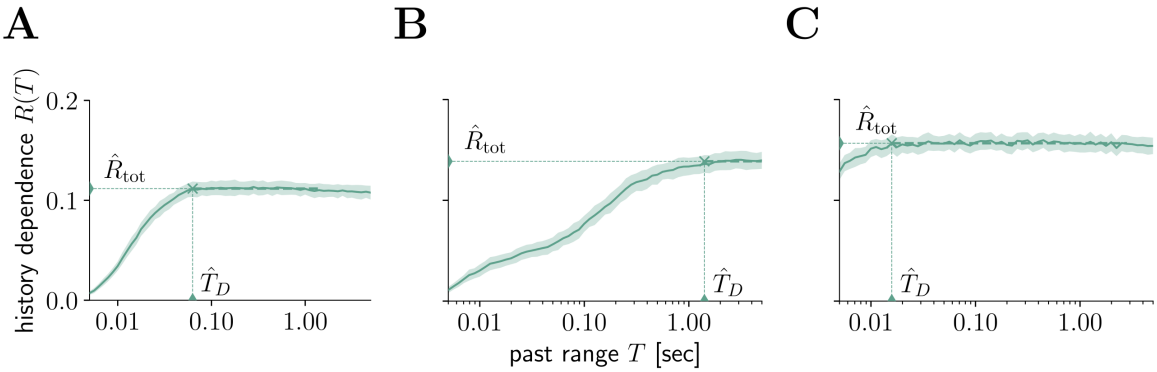

Fig 6. Distinct signatures of history dependence for different neurons within mouse primary visual cortex. Embedding-optimized estimates of $R(T)$ reveal distinct signatures of history dependence for different neurons within a single recorded system (mouse primary visual cortex). In particular, neurons have similar total history dependence $\hat{R}_{\text{tot}}$, but differ vastly in the estimated temporal depth $\hat{T}_D$. Shaded areas indicate standard deviation obtained by bootstrapping, and the dashed line indicates past ranges over which estimates $\hat{R}(T)$ were averaged to compute $\hat{R}_{\text{tot}}$. Estimates were computed with the Shuffling estimator and $d_{\max} = 5$.

$T \approx 200$ ms in EC (Fig 4B), possibly reflecting phase information in the theta cycles [32, 33].

Overall, this demonstrates that embedding optimization is powerful enough to reveal clear differences in history dependence between neurons of different recorded systems, but also between neurons within the same system. Even more importantly, because neurons are so different, ad hoc embedding schemes with a fixed number of bins or fixed bin width will miss considerable history dependence.

Discussion

To estimate history dependence in neural spiking activity, we developed a method where the embedding of past spiking is optimized for each individual neuron. Thereby, it can estimate a maximum of history dependence, given what is possible for the limited amount of data. We found that embedding optimization is a robust and flexible tool to estimate history dependence in neural spike trains with vastly different spiking statistics, where ad hoc embedding strategies would estimate substantially less history dependence.

In the following, we contrast history dependence R with pairwise measures such as the auto-correlation. We then discuss how one can relate estimated history dependence to neural coding and information processing at the example data sets analyzed in this paper.

Why quantify history dependence and not the auto-correlation or auto-information? First, history dependence $R(T)$ captures an important footprint of neural coding that is not captured by pairwise dependency measures such as the auto-correlation [38] or the auto- or delayed mutual information [39]. History dependence $R(T)$ quantifies statistical dependencies between current spiking and past spiking in the entire past range T. From a coding perspective, it gives the redundancy of neural spiking with all past spikes in the past range T, or how much past information in T is integrated when emitting a spike. In contrast, auto-correlation or auto-information quantify the statistical dependency of neural spiking onto a single past bin, independent of all other past bins. Thereby, these measures neglect dependencies that only show in the context of other bins. Moreover, they miscount dependencies that vanish once spiking activity in more recent past is taken into consideration.

Second, quantifying history dependence yields the temporal depth, which provides an intrinsic timescale of single neuron spiking with respect to all linear, non-linear and higher-order dependencies.

Previously, the intrinsic timescale was quantified by the autocorrelation time [8, 40], which takes into account linear and pairwise history dependence. The autocorrelation time is of interest, because it is related to recurrent connection strength and reverberations of activity in a simple model of neural activity propagation [41, 42, 43]. For any deviations from the simple model, however, the autocorrelation might not decay exponentially, and history dependence might be dominated by non-linear contributions. The temporal depth of history dependence, in contrast, remains well-defined and considers linear as well as non-linear contributions alike.

A downside of quantifying the history dependence R is that its estimation requires more data. To make best use of the limited data, we here devised an embedding optimization approach that allows to find an efficient representation of past spiking for the estimation of history dependence. Nonetheless, we found empirically that a minimum of 10 minutes of recorded spiking activity are advisable to allow a meaningful quantification of history dependence and its temporal depth (supplementary Fig S2).

Differences in history dependence and temporal depth between data sets agree with ideas from neural coding and hierarchical information processing. First, we found that the estimated total history dependence $\hat{R}_{\text{tot}}$ clearly differs among the experimental data sets. Notably, $\hat{R}_{\text{tot}}$ was low for recordings of early visual processing areas such as retina and primary visual cortex, which is in line with the theory of efficient coding [1, 44] and neural adaptation for temporal whitening as observed in experiments [3, 45]. In contrast, $\hat{R}_{\text{tot}}$ was high for neurons in dorsal hippocampus (layer CA1) and cortical culture. In CA1, the original study [33] found that the temporal structure of neural activity within the temporal windows set by the theta cycles was beyond of what one would expect from integration of feed-forward excitatory inputs. The authors concluded that this could be due to local circuit computations. The high values of $\hat{R}_{\text{tot}}$ support this idea, and suggest that local circuit computations could serve the integration of past information, either for the formation of a path integration–based neural map [46], or to recognize statistical structure for associative learning [7]. In cortical culture, neurons are exclusively driven by recurrent input and exhibit strong bursts in the population activity [47]. This leads to strong history dependence also at the single neuron level.

To summarize, history dependence was low for early sensory processing and high for high level processing or past dependencies that are induced by strong recurrent feedback in a neural network. We thus conclude that estimated total history dependence $\hat{R}_{\text{tot}}$ does indeed provide a footprint of neural coding and information processing.

Second, we observed that the temporal depth T_D of history dependence increases from retina ($\approx 70\,\text{ms}$) to primary visual cortex ($\approx 160\,\text{ms}$) to EC ($\approx 450\,\text{ms}$), in agreement with the idea of a temporal hierarchy in neural information processing [9]. These results agree with similar results obtained for the autocorrelation time of spontaneous activity [8]. Our results indicate that the hierarchy of intrinsic timescales is also reflected in the history dependence of single neurons measured by the mutual information.

Conclusion. Embedding optimization enables to estimate history dependence in a diversity of spiking neural systems, both in terms of the magnitude as well as the temporal depth. The approach could be used in future experimental studies to quantify history dependence across a diversity of brain areas, e.g. using the novel neuropixel probe, or even across cortical layers within a single area. To this end we provide a toolbox for Python3 [29] and practical guidelines in the Materials and methods section. These analyses might yield a more complete picture of hierarchical processing in terms of the timescale *and* a footprint of information processing and coding principles, i.e. information integration versus redundancy reduction.

Materials and methods

In this section, we provide all mathematical details required to reproduce the results of this paper. We first provide the basic definitions of history dependence, the past embedding as well as the total history dependence and its temporal depth. We then describe the embedding optimization approach that is used to estimate history dependence from neural spike recordings, and provide a description of the workflow. Next, we delineate the estimators of history dependence considered in this paper, and present the novel Bayesian bias criterion. Finally, we provide details on the benchmark model and how we approximated its history dependence for given past range and embedding parameters. All code for Python3 that was used to analyze the data and to generate the figures is available online at `https://github.com/Priesemann-Group/historydependence`.

Glossary

Terms

- *Past embedding*: discrete, reduced representation of past spiking through temporal binning
- *Past-embedding optimization*: Optimization of temporal binning for better estimation of history dependence
- *Embedding-optimized estimate*: Estimate of history dependence for optimized embedding

Abbreviations

- *GLM*: generalized linear model
- *ML*: Maximum likelihood
- *BBC*: Bayesian bias criterion
- *Shuffling*: Shuffling estimator based on a bias correction for the ML estimator

Symbols

- Δt: bin size of the time bin for current spiking
- T: past range of the past embedding
- $[t-T, t)$: embedded past window
- d: embedding dimension or number of bins
- κ: scaling exponent for exponential embedding
- T_{rec}: recording length
- $N = (T_{\text{rec}} - T)/\Delta t$: number of measurements, i.e. number of observed joint events of current and past spiking
- X: random variable with binary outcomes $x \in [0, 1]$, which indicate the presence of a spike in a time bin Δt
- $\boldsymbol{X}^{-T}$: random variable whose outcomes are binary sequences $\boldsymbol{x}^{-T} \in \{0, 1\}^d$, which represent past spiking activity in a past range T

Information theoretic quantities

- $H(\text{spiking}) \equiv H(X)$: average spiking information
- $H(\text{spiking}|\text{past}) \equiv H(X|\boldsymbol{X}^{-T})$: average spiking information for given past spiking in a past range T
- $I(\text{spiking}; \text{past}) \equiv I(X; \boldsymbol{X}^{-T})$: mutual information between current spiking and past spiking in a past range T
- $R(T) \equiv I(X; \boldsymbol{X}^{-T})/H(X)$: history dependence for given past range T
- $R(T, d, \kappa) \equiv I(X; \boldsymbol{X}^{-T}_{d,\kappa})/H(X)$: history dependence for given past range T and past embedding d, κ
- $R_{\text{tot}} \equiv \lim_{T\to\infty} R(T)$: total history dependence
- T_D: temporal depth, i.e. minimal T for which $R(T) = R_{\text{tot}}$

Estimated quantities

- $\hat{R}(T, d, \kappa)$: estimated history dependence for given past range T and past embedding d, κ
- $\hat{R}(T)$: embedding-optimized estimate of $R(T)$ for optimal embedding parameters d^*, κ^*
- $\hat{T}_D$: estimated temporal depth, i.e. past range T for which $\hat{R}(T)$ saturates within errorbars
- $\hat{R}_{\text{tot}}$: estimated total history dependence, i.e. average $\hat{R}(T)$ for $T \in [\hat{T}_D, T_{\max}]$

Basic definitions

Definition of history dependence. We estimate history dependence $R(T)$ as the mutual information $I(X, \boldsymbol{X}^{-T})$ between present and past spiking X and $\boldsymbol{X}^{-T}$, normalized by the binary Shannon

information of spiking $H(X)$, i.e.

$$R(T) \equiv \frac{I(X, \boldsymbol{X}^{-T})}{H(X)} = 1 - \frac{H(X|\boldsymbol{X}^{-T})}{H(X)}. \tag{1}$$

Under the assumption of stationarity and ergodicity the mutual information can be computed either as the average over the stationary distribution $p(x, \boldsymbol{x}^{-T})$, or the time average [13, 48], i.e.

$$\begin{aligned} I(X, \boldsymbol{X}^{-T}) &= H(X) - H(X|\boldsymbol{X}^{-T}) && (2)\\ &= \sum_{x\in\{0,1\}} p(x) \log_2 \frac{1}{p(x)} - \sum_{\boldsymbol{x}^{-T}\in\{0,1\}^d} p(x, \boldsymbol{x}^{-T}) \log_2 \frac{1}{p(x|\boldsymbol{x}^{-T})} && (3)\\ &= \sum_{x\in\{0,1\}} \sum_{\boldsymbol{x}^{-T}\in\{0,1\}^d} p(x, \boldsymbol{x}^{-T}) \log_2 \frac{p(x|\boldsymbol{x}^{-T})}{p(x)} && (4)\\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \log_2 \frac{p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T})}{p(x_{t_n})}. && (5) \end{aligned}$$

Here, $x_{t_n} \in \{0, 1\}$ indicates the presence of a spike in a small interval $[t_n, t_n + \Delta t]$ with $\Delta t = 5\,\text{ms}$ throughout the paper, and $\boldsymbol{x}_{t_n}^{-T}$ encodes the spiking history in a time window $[t_n - T, t_n)$ at times $t_n = n\Delta t$ that are shifted by Δt.

Past embedding. Here, we encode the spiking history in a finite time window $[t - T, t)$ as a binary sequence $\boldsymbol{x}_t^{-T} = (x_{t,i}^{-T})_{i=1}^d$ of binary spike counts $x_{t,i}^{-T} \in \{0, 1\}$ in d past bins (Fig 2). When more than one spike can occur in a single bin, $x_{t,i}^{-T} = 1$ is chosen for spike counts larger than the median activity in the ith bin. This type of temporal binning is more generally referred to as *past embedding*. It is formally defined as a mapping

$$\Gamma_T(\theta) : \mathcal{F}_T \to S^d \tag{6}$$

from the set of all possible spiking histories $\mathcal{F}_T = \sigma(\mathcal{X}_\tau : \tau \in [t - T, t))$, i.e. the sigma algebra generated by the point process $\mathcal{X}$ (neural spiking) in the time interval $[t - T, t)$, to the set of d-dimensional binary sequences S^d. We can drop the dependence on the time t because we assume stationarity of the point process. Here, T is the embedded *past range*, d the *embedding dimension*, and θ denotes all the embedding parameters that govern the mapping, i.e. $\theta = (d, ...)$. The resulting binary sequence at time t for given embedding θ and past range T will be denoted by $\boldsymbol{x}_{t,\theta}^{-T}$. In this paper, we consider the following two embeddings for the estimation of history dependence.

Uniform embedding. If all bins have the same bin width $\tau = T/d$, the embedding is called *uniform*. The main drawback of the uniform embedding is that higher past ranges T enforce a uniform decrease in resolution τ when d is fixed.

Exponential embedding. One can generalize the uniform embedding by letting bin widths increase exponentially with bin index $j = 1, ..., d$ according to $\tau_j = \tau 10^{(j-1)\kappa}$. Here, τ gives the bin size of the first past bin, and is uniquely determined when T, d and κ are specified. Note that $\kappa = 0$ yields a uniform embedding, whereas $\kappa > 0$ decreases resolution on distant past spikes. For fixed embedding dimension d and past range T, this allows to retain a higher resolution on spikes in the more recent past.

Sufficient embedding. Ideally, the past embedding preserves all the information that the spiking history in the past range T has about the present spiking dynamics. In that case, no additional past information has an influence on the probability for x_t once the embedded spiking history $\boldsymbol{x}_{t,\theta}^{-T}$ is given, i.e.

$$p(x_t|\boldsymbol{x}_{t,\theta}^{-T}, \boldsymbol{x}_{t,\nu}^{-T}) = p(x_t|\boldsymbol{x}_{t,\theta}^{-T}) \tag{7}$$

for any other past embedding $\boldsymbol{x}_{t,\nu}^{-T}$. If Eq (7) holds for all times t, the embedding $\Gamma_T(\theta)$ is called a *sufficient* embedding. For the remainder of this paper, the sequences of sufficient embeddings are denoted by $\boldsymbol{x}_t^{-T}$.

Insufficient embeddings cause underestimation of history dependence. The past embedding is essential when inferring history dependence from recordings, because an insufficient embedding causes underestimation of history dependence. To show this, we note that for any embedding parameters θ and past range T the Kullback-Leibler divergence between the spiking probability for the sufficient embedding $p(x_t|\boldsymbol{x}_t^{-T})$ and $p(x_t|\boldsymbol{x}_{t,\theta}^{-T})$ cannot be negative [49], i.e.

$$D_{KL}\left[p(x_t|\boldsymbol{x}_t^{-T})||p(x_t|\boldsymbol{x}_{t,\theta}^{-T})\right] = \sum_{x_t\in\{0,1\}} p(x_t|\boldsymbol{x}_t^{-T})\log_2\frac{p(x_t|\boldsymbol{x}_t^{-T})}{p(x_t|\boldsymbol{x}_{t,\theta}^{-T})} \geq 0, \tag{8}$$

with equality *iff* $p(x_t|\boldsymbol{x}_{t,\theta}^{-T}) = p(x_t|\boldsymbol{x}_t^{-T})$. By taking the average over all times t_n, we arrive at

$$\begin{aligned}
0 &\leq \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\sum_{x_{t_n}\in\{0,1\}} p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T})\log_2\frac{p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T})}{p(x_{t_n}|\boldsymbol{x}_{t_n,\theta}^{-T})} && (9)\\
&= \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\sum_{x_{t_n}\in\{0,1\}} p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T},\boldsymbol{x}_{t_n,\theta}^{-T})\log_2\frac{1}{p(x_{t_n}|\boldsymbol{x}_{t_n,\theta}^{-T})} && (10)\\
&\quad - \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\sum_{x_{t_n}\in\{0,1\}} p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T})\log_2\frac{1}{p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T})} && (11)\\
&= H(X|\boldsymbol{X}_\theta^{-T}) - H(X|\boldsymbol{X}^{-T}), && (12)
\end{aligned}$$

where the last step follows from stationarity and ergodicity and marginalizing out $\boldsymbol{x}_{t_n}^{-T}$ in the first term. From here, it follows that one always underestimates the history dependence in neural spiking, as long as the embedding is not sufficient, i.e.

$$R(T,\theta) \equiv 1 - \frac{H(X|\boldsymbol{X}_\theta^{-T})}{H(X)} \leq 1 - \frac{H(X|\boldsymbol{X}^{-T})}{H(X)} = R(T). \tag{13}$$

Total history dependence and temporal depth. In this paper we quantify history dependence $R(T)$ in dependence of the past range T. This allows us to characterize history dependence not only in terms of the *total history dependence* R_{tot}, but also the *temporal depth* T_D. We defined the total history dependence as the limit for an infinite past range

$$R_{\text{tot}} \equiv \lim_{T\to\infty} R(T), \tag{14}$$

and quantifies all dependencies of neural spiking on its own spiking history. The temporal depth we defined as the minimal past range T for which the history dependence is equal to the total history dependence, i.e.

$$T_D \equiv \min T|_{R(T)=R_{\text{tot}}}. \tag{15}$$

The temporal depth T_D gives the past range over which spiking depends on its own history.

Estimation of history dependence using past-embedding optimization

The past embedding is crucial in determining how much history dependence we can capture, since an insufficient embedding θ leads to an underestimation of the history dependence $R(T) \geq R(T,\theta)$. In order to capture as much history dependence as possible, the embedding θ should be chosen to maximize the estimated history dependence $R(T,\theta)$. Since the history dependence has to be estimated from data, we formulate the following embedding optimization procedure in terms of the estimated history dependence $\hat{R}(T,\theta)$.

Embedding optimization. For given T, find the optimal embedding θ^* that maximizes the estimated history dependence

$$\theta^* = \arg\max_{\theta} \hat{R}(T, \theta). \tag{16}$$

This yields an *embedding-optimized* estimate $\hat{R}(T) = \hat{R}(T, \theta^*)$ of the true history dependence $R(T)$.

Requirements. Embedding optimization can only give sensible results if the optimized estimates $\hat{R}(T, \theta)$ are guaranteed to be unbiased or a lower bound to the true $R(T, \theta)$. Otherwise, embeddings will be chosen that strongly overestimate history dependence. In this paper, we therefore use two estimators, BBC and Shuffling, the former of which is designed to be unbiased, and the latter a lower bound to the true $R(T, \theta)$ (see below). In addition, embedding optimization works only if the estimation variance is sufficiently small. Otherwise, maximizing over variable estimates can lead to a mild overestimation. We found for a benchmark model that this overestimation was negligibly small for a recording length of 90 minutes for a model neuron with a 4 Hz average firing rate (supplementary Fig S1). For smaller recording lengths, potential overfitting can be avoided by cross-validation, i.e. optimizing embeddings on one half of the recording and computing embedding-optimized estimates on the other half.

Implementation. For the optimization, we compute estimates $\hat{R}(T, d, \kappa)$ for a range of embedding dimensions $d \in [1, 2, ..., d_{\max}]$ and scaling parameter $\kappa = [0, ..., \kappa_{\max}]$. For each T, we then choose the optimal parameter combination d^*, κ^* for each T that maximizes the estimated history dependence $\hat{R}(T, d, \kappa)$, and use $\hat{R}(T, d^*, \kappa^*)$ as the best estimate of $R(T)$.

Estimation of temporal depth and total history dependence. Using the embedding-optimized estimates $\hat{R}(T)$, the temporal depth T_D is estimated as the minimum past range T for which $\hat{R}(T)$ lies within one standard deviation of the maximum estimated history dependence, i.e.

$$\hat{T}_D \equiv \min T|_{R(T) \geq \hat{R}_{\max} - \sigma_{\hat{R}_{\max}}}, \tag{17}$$

with

$$\hat{R}_{\max} = \max_T \hat{R}(T). \tag{18}$$

The standard deviation $\sigma_{\hat{R}_{\max}}$ was estimated by bootstrapping (see below). Taking the standard deviation into account makes estimates of the temporal depth more robust to statistical fluctuations in estimates of the history dependence $\hat{R}(T)$. The total history dependence was estimated by averaging $\hat{R}(T)$ over past ranges $T \in [\hat{T}_D, T_{\max}]$ that were larger or equal to the temporal depth, but not larger than $T_{\max}$. The maximum past range was chosen as the highest past range for which $\hat{R}(T)$ lies within standard error of the maximum estimated history dependence, i.e.

$$T_{\max} \equiv \max T|_{R(T) \geq \hat{R}_{\max} - \sigma_{\hat{R}_{\max}}}. \tag{19}$$

This avoids averaging over estimates that are systematically underestimated because of limited resolution for high past ranges.

Workflow. The estimation workflow using embedding optimization can be summarized by the following sequence of steps (Fig 1):

1) Define a set of embedding parameters d, κ for fixed past range T.
2) For each embedding d, κ, record sequences of current and past spiking $x_{t_n}, \boldsymbol{x}_{t_n,\theta}^{-T}$ for all time steps t_n in the recording.
3) Use the frequencies of the recorded sequences to estimate history dependence for each embedding.
4) Apply regularization such that all estimates are unbiased or lower bounds to the true history dependence.

5) Select the optimal embedding to obtain an embedding-optimized estimate $\hat{R}(T)$.

6) Repeat the estimation for a set of past ranges T to obtain estimates of the temporal depth $\hat{T}_D$ and the total history dependence $\hat{R}_{\text{tot}}$.

Different estimators of history dependence

To estimate $R(T,\theta)$, one has to estimate the binary entropy of spiking $H(X)$ in a small time bin Δt, and the conditional entropy $H(X|\boldsymbol{X}_\theta^{-T})$ from data. The estimation of the binary entropy only requires the average firing probability $p(x{=}1) = r\Delta t$ with

$$\hat{H}(X) = -r\Delta t \log_2 r\Delta t - (1 - r\Delta t)\log_2(1 - r\Delta t), \tag{20}$$

which can be estimated with high accuracy from the estimated average firing rate r even for short recordings. The conditional entropy $H(X|\boldsymbol{X}_\theta^{-T})$, on the other hand, is much more difficult to estimate. In this paper, we focus on a non-parametric approach that estimates

$$H(X|\boldsymbol{X}_\theta^{-T}) = H(X, \boldsymbol{X}_\theta^{-T}) - H(\boldsymbol{X}_\theta^{-T}) \tag{21}$$

by a non-parametric estimation of the entropies $H(\boldsymbol{X}_\theta^{-T})$ and $H(X, \boldsymbol{X}_\theta^{-T})$.
The estimation of entropy from data is a well-established problem, and we can make use of previously developed entropy estimation techniques for the estimation of history dependence. We here write out the estimation of the entropy term for joint sequences of present and past spiking $H(X, \boldsymbol{X}_\theta^{-T})$, which is the highest dimensional term and thus the hardest to estimate. Estimation for the marginal entropy $H(\boldsymbol{X}_\theta^{-T})$ is completely analogous.

Computing the entropy requires knowing the statistical uncertainty and thus the probabilities for all possible joint sequences. In the following we will write probabilities as a vector $\boldsymbol{\pi} = (\pi_k)_{k=1}^K$, where $\pi_k \equiv p\left((x, \boldsymbol{x}_\theta^{-T}){=}a_k\right)$ are the probabilities for the $K = 2^{d+1}$ possible joint spike pattern $a_k \in \{0,1\}^{d+1}$. The entropy $H(X, \boldsymbol{X}_\theta^{-T})$ then reads

$$H(X, \boldsymbol{X}_\theta^{-T}) = H(\boldsymbol{\pi}) = -\sum_{k=1}^{K} \pi_k \log_2 \pi_k. \tag{22}$$

Once we are able to estimate the probability distribution $\boldsymbol{\pi}$, we are able to estimate the entropy. In a non-parametric approach, the probabilities $\boldsymbol{\pi} = (\pi_k)_{k=1}^K$ are directly inferred from counts $\boldsymbol{n} = (n_k)_{k=1}^K$ of different spike sequences a_k within the spike recording. Each timestep $[t_n, t_n + \Delta t]$ provides a sample of present spiking x_{t_n} and its history $\boldsymbol{x}_{t_n,\theta}^{-T}$, such that a recording of length T_{rec} provides $N = (T_{\text{rec}} - T)/\Delta t$ data points.

Maximum likelihood estimation. Most commonly, probabilities of spike sequences a_k are then estimated as the relative frequencies $\hat{\pi}_k = n_k/N$ of their occurrence in the observed data. It is the maximum likelihood (ML) estimator of $\boldsymbol{\pi}$ for the multinomial likelihood

$$p(\boldsymbol{n}|\boldsymbol{\pi}) \propto \prod_{k=1}^{K} \pi_k^{n_k}. \tag{23}$$

Plugging the estimates $\hat{\pi}_k$ into the definition of entropy results in the 'ML' estimator of the entropy

$$\hat{H}_{\text{ML}}(X, \boldsymbol{X}_\theta^{-T}) = -\sum_{k=1}^{K} \frac{n_k}{N} \log_2 \frac{n_k}{N} \tag{24}$$

or history dependence

$$\hat{R}_{\text{ML}}(T,\theta) = 1 - \frac{\hat{H}_{\text{ML}}(X, \boldsymbol{X}_\theta^{-T}) - \hat{H}_{\text{ML}}(\boldsymbol{X}_\theta^{-T})}{\hat{H}(X)}. \tag{25}$$

The ML estimator has the right asymptotic properties [20, 50], but is known to underestimate the entropy severely when data is limited [20, 51]. This is because all probability mass is assumed to be concentrated on the *observed* outcomes. A more concentrated probability distribution results in a smaller entropy, in particular if many outcomes have not been observed. This results in a systematic underestimation or negative bias

$$\text{bias}\left[\hat{H}_{\text{ML}}(X, \boldsymbol{X}_\theta^{-T})\right] \leq 0. \tag{26}$$

The negative bias in the entropy, which is largest for the highest-dimensional joint entropy $\hat{H}_{\text{ML}}(X, \boldsymbol{X}_\theta^{-T})$, then typically leads to severe overestimation of the mutual information and history dependence [19, 52]. Because of this severe overestimation, we cannot use the ML estimator for embedding optimization.

Bayesian Nemenman-Shafee-Bialek (NSB) estimator. In a Bayesian framework, the entropy is estimated as the posterior mean or minimum mean square error (MMSE)

$$\hat{H}_{\text{MMSE}}(\boldsymbol{n}) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{n}) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) \frac{p(\boldsymbol{n}|\boldsymbol{\pi})p(\boldsymbol{\pi})}{\int d\boldsymbol{\pi}' p(\boldsymbol{n}|\boldsymbol{\pi}')p(\boldsymbol{\pi}')}. \tag{27}$$

The posterior mean is the mean of the entropy with respect to the posterior distribution on the probability vector $\boldsymbol{\pi}$ given the observed frequencies of spike sequences $\boldsymbol{n}$

$$p(\boldsymbol{\pi}|\boldsymbol{n}) = \frac{p(\boldsymbol{n}|\boldsymbol{\pi})p(\boldsymbol{\pi})}{\int d\boldsymbol{\pi}' p(\boldsymbol{n}|\boldsymbol{\pi}')p(\boldsymbol{\pi}')}. \tag{28}$$

The probability for i.i.d. observations $\boldsymbol{n}$ from an underlying distribution $\boldsymbol{\pi}$ is given by the multinomial distribution in Eq (23).

If the prior $p(\boldsymbol{\pi})$ is a conjugate prior to the multinomial likelihood, then the high dimensional integral of Eq (27) can be evaluated analytically [24]. This is true for a class of priors called Dirichlet priors, and in particular for symmetric Dirichlet priors

$$p(\boldsymbol{\pi}|\beta) \propto \prod_{k=1}^{K} \pi_k^{\beta-1}. \tag{29}$$

The prior $p(\boldsymbol{\pi}|\beta)$ gives every outcome the same a priori weight, but controls the weight $\beta > 0$ of uniform prior pseudo-counts. A $\beta = 1$ corresponds to a flat prior on all probability distributions $\boldsymbol{\pi}$, whereas $\beta \to 0$ gives maximum likelihood estimation (no prior pseudo-count).

It has been shown that the choice of β is highly informative with respect to the entropy, in particular when the number of outcomes K becomes large [53]. This is because the a priori variance of the entropy vanishes for $K \to \infty$, such that for any $\boldsymbol{\pi} \sim p(\boldsymbol{\pi}|\beta)$ the entropy $H(\boldsymbol{\pi})$ is very close to the a priori expected entropy

$$\xi(\beta) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) p(\boldsymbol{\pi}|\beta) = \psi_0(K\beta + 1) - \psi_0(\beta + 1), \tag{30}$$

where $\psi_m(z) = \partial_z^{m+1} \log \Gamma(z)$ are the polygamma functions. In addition, a lot of data is required to counter-balance this a priori expectation. The reason is the prior adds pseudo-counts on every outcome, i.e. it assumes that every outcome has been observed β times prior to inference. In order to influence a prior that constitutes K pseudo-counts, one needs at least $N > K$ samples, with more data required the sparser the true underlying distribution. Therefore, an estimator of the entropy for little data and fixed concentration parameter β is highly biased towards the a priori expected entropy $\xi(\beta)$.

Nemenman et al. [25] exploited the tight link between concentration parameter β and the a priori expected entropy to derive a mixture prior

$$p_{NSB}(\boldsymbol{\pi}) \propto \int d\beta \left|\frac{\partial \xi}{\partial \beta}\right| p(\boldsymbol{\pi}|\beta), \tag{31}$$

$$\frac{\partial \xi}{\partial \beta} = K\psi_1(K\beta + 1) - \psi_1(\beta + 1), \tag{32}$$

that weights Dirichlet priors to be flat with respect to the expected entropy $\xi(\beta)$. Since the variance of this expectation vanishes for $K \gg 1$ [53], for high K the prior is also approximately flat with respect to the entropy, i.e. $H(\boldsymbol{\pi}) \sim \mathcal{U}(0, \log_2 K)$ for $\boldsymbol{\pi} \sim p_{NSB}(\boldsymbol{\pi})$. The resulting MMSE estimator for the entropy is referred to as the NSB estimator

$$\hat{H}_{NSB}(\boldsymbol{n}) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) \frac{p(\boldsymbol{n}|\boldsymbol{\pi}) p_{NSB}(\boldsymbol{\pi})}{\int d\boldsymbol{\pi}' p(\boldsymbol{n}|\boldsymbol{\pi}') p_{NSB}(\boldsymbol{\pi}')} \tag{33}$$

$$= \frac{\int d\beta \frac{d\xi}{d\beta}(\beta) \hat{H}(\beta) \rho(\beta, \boldsymbol{n})}{\int d\beta' \frac{d\xi}{d\beta}(\beta') \rho(\beta', \boldsymbol{n})}. \tag{34}$$

Here, $\rho(\beta, \boldsymbol{n})$ is proportional to the evidence for given concentration parameter

$$\rho(\beta, \boldsymbol{n}) := \frac{\Gamma(K\beta)}{\Gamma(N + K\beta)} \prod_{i=1}^{K} \frac{\Gamma(n_i + \beta)}{\Gamma(\beta)} \tag{35}$$

$$\propto \int d\boldsymbol{\pi}\, p(\boldsymbol{n}|\boldsymbol{\pi})\, p(\boldsymbol{\pi}|\beta) = p(\boldsymbol{n}|\beta), \tag{36}$$

where $\Gamma(x)$ is the gamma function. The posterior mean of the entropy for given concentration parameter is

$$\hat{H}(\beta) = \sum_{i=1}^{K} \frac{n_i + \beta}{N + K\beta} [\psi_0(N + K\beta + 1) - \psi_0(n_i + \beta + 1)]. \tag{37}$$

From the Bayesian entropy estimate, we obtain an NSB estimator for history dependence

$$\hat{R}_{\mathrm{NSB}}(T, \theta) = 1 - \frac{\hat{H}_{\mathrm{NSB}}(X, \boldsymbol{X}_\theta^{-T}) - \hat{H}_{\mathrm{NSB}}(\boldsymbol{X}_\theta^{-T})}{\hat{H}(X)}. \tag{38}$$

where the marginal and joint entropies are estimated individually using the NSB method.

To compute the NSB entropy estimator, one has to perform a one-dimensional integral over all possible concentration parameters β. This is crucial to be unbiased with respect to the entropy. An implementation of the NSB estimator for Python3 is published alongside the paper with our toolbox [29]. To compute the integral, we use a Gaussian approximation around the maximum a posteriori β^* to define sensible integration bounds when the likelihood is highly peaked, as proposed in [26].

Bayesian bias criterion. The goal of the Bayesian bias criterion (BBC) is to indicate when estimates of history dependence are potentially biased. It might indicate bias even when estimates are unbiased, but the opposite should never be true.

To indicate a potential estimation bias, the BBC compares ML and BBC estimates of the history dependence. ML estimates are biased when too few joint sequences have been observed, such that the probability for unobserved or undersampled joint outcomes is underestimated. To counterbalance this effect, the NSB estimate adds β pseudo-counts to every outcome, and then infers β with an uninformative prior. For the BBC, we turn the idea around: when the assumption of no pseudo-counts (ML) versus a posterior belief on non-zero pseudo-counts (NSB) yield different estimates of history dependence, then too few sequences have been observed and estimates are potentially biased. This motivates the following definition of the BBC.

The NSB estimator $R_{\mathrm{NSB}}(T, \theta)$ is biased with tolerance $p > 0$, if

$$|\hat{R}_{\mathrm{NSB}}(T, \theta) - \hat{R}_{\mathrm{ML}}(T, \theta)| > p \cdot \hat{R}_{\mathrm{NSB}}(T, \theta). \tag{39}$$

Similarly, we define the BBC estimator

$$\hat{R}_{\mathrm{BBC}}(T, \theta) \equiv \begin{cases} \hat{R}_{\mathrm{NSB}}(T, \theta) & \text{if} \quad \hat{R}_{\mathrm{NSB}}(T, \theta) - \hat{R}_{\mathrm{ML}}(T, theta) \leq p \cdot \hat{R}_{\mathrm{NSB}}(T, \theta), \\ 0 & \text{otherwise.} \end{cases} \tag{40}$$

This estimator is designed to be unbiased, and can thus can be used for embedding optimization in Eq (16). We use the NSB estimator for $R(T,\theta)$ instead of the ML estimator, because it is generally less biased. A tolerance $p > 0$ accounts for this, and accepts NSB estimates when there is only a small difference between the estimates. The bound for the difference is multiplied by $\hat{R}_{\mathrm{NSB}}(T,\theta)$, because this provides the scale on which one should be sensitive to estimation bias. We found that a tolerance of $p = 0.05$ was small enough to avoid overestimation by BBC estimates on the benchmark model (Fig 3 and supplementary Fig S1).

Shuffling estimator. The Shuffling estimator was originally proposed in [23] to reduce the sampling bias of the ML mutual information estimator. It has the desirable property that it is negatively biased in leading order of the inverse number of samples. Because of this property, Shuffling estimates can safely be maximized during embedding optimization without the risk of overestimation. Here, we therefore propose to use the Shuffling estimator for embedding-optimized estimation of history dependence.

The idea behind the Shuffling estimator is to rewrite the ML estimator of history dependence as

$$\hat{R}_{\mathrm{ML}}(T,\theta) = \frac{1}{\hat{H}(X)} \left(\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}) - \hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X) \right) \tag{41}$$

and to correct for bias in the entropy estimate $\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)$. Since X is well sampled and thus $\hat{H}(X)$ is unbiased, and the bias of the ML entropy estimator is always negative [20, 51], we know that

$$\begin{aligned} \mathrm{Bias}[\hat{R}_{\mathrm{ML}}(T,\theta)] &= \mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T})] - \mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)] && (42) \\ &\leq -\mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)]. && (43) \end{aligned}$$

Therefore, if we find a correction term of the magnitude of $\mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)]$, we can turn the bias in the estimate of the history dependence from positive to negative, thus obtaining an estimator that is a lower bound of the true history dependence. This can be achieved by subtracting a lower bound of the estimation bias $\mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)]$ from $\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)$.

In the following, we describe how [23] obtain a lower bound of the bias in the conditional entropy $\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)$ by computing the estimation bias for shuffled surrogate data.
Surrogate data are created by shuffling recorded spike sequences such that statistical dependencies between past bins are eliminated. This is achieved by taking all past sequences that were followed by a spike, and permuting past observations of the same bin index j. The same is repeated for all past sequences that were followed by no spike. The underlying probability distribution can then be computed as

$$p_{\mathrm{sh}}(\boldsymbol{x}_\theta^{-T}|x) = \prod_{j=1}^{d} p(x_{\theta,j}^{-T}|x), \tag{44}$$

and the corresponding entropy is

$$H(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X) = \sum_{j=1}^{d} H(X_j^{-T}|X). \tag{45}$$

The pairwise probabilities $p(x_{\theta,j}^{-T}|x)$ are well sampled, and thus each conditional entropy in the sum can be estimated with high precision. This way, the true conditional entropy $H(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X)$ for the shuffled surrogate data can be computed and compared to the ML estimate $\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X)$ on the shuffled data. The difference between the two

$$\Delta\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X)] \equiv \hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X) - H(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X) \tag{46}$$

yields a correction term that is on average equal to the bias of the ML estimator on the shuffled data.

Importantly, the bias of the ML estimator on the shuffled data is in leading order more negative than on the original data. To see this, we consider an expansion of the bias on the conditional entropy in inverse powers of the sample size N [19, 52]

$$\mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)] = -\frac{1}{2N\ln 2}\sum_{x\in\{0,1\}}\left(\tilde{K}(x)-1\right)+\mathcal{O}\left(\frac{1}{N^2}\right). \tag{47}$$

Here, $\tilde{K}(x)$ denotes the number of past sequences with nonzero probability $p(\boldsymbol{x}_\theta^{-T}{=}a_k|x)>0$ of being observed when followed by a spike ($x=1$) or no spike ($x=0$), respectively. Notably, the bias is negative in leading order, and depends only on the number of possible sequences $\tilde{K}(x)$. For the shuffled surrogate data, we know that $p_{\mathrm{sh}}(\boldsymbol{x}_\theta^{-T}{=}a_k|x)=0$ implies $p(\boldsymbol{x}_\theta^{-T}{=}a_k|x)=0$, but Shuffling may lead to novel sequences that have zero probability otherwise. Hence the number of possible sequences under Shuffling can only increase, i.e. $\tilde{K}_{\mathrm{sh}}(x)\geq\tilde{K}(x)$, and thus the bias of the ML estimator under Shuffling to first order is always more negative than for the original data

$$\mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X)] \lesssim \mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)]. \tag{48}$$

Terms that could render it higher are of order $\mathcal{O}(N^{-2})$ and higher and are assumed to have no practical relevance.

This motivates the following definition of the Shuffling estimator: Compute the difference between the ML estimator on the shuffled and original data to yield a bias-corrected Shuffling estimate

$$\hat{H}_{\mathrm{ML,sh}}(\boldsymbol{X}_\theta^{-T}|X) \equiv \hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X) - \Delta\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X), \tag{49}$$

and use this to estimate history dependence

$$\hat{R}_{\mathrm{Shuffling}}(T,\theta) \equiv \frac{1}{\hat{H}(X)}\left(\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}) - \hat{H}_{\mathrm{ML,sh}}(\boldsymbol{X}_\theta^{-T}|X)\right). \tag{50}$$

Because of Eq (43) and Eq (48), we know that this estimator is negatively biased in leading order

$$\hat{R}_{\mathrm{Shuffling}}(T,\theta) \lesssim 0 \tag{51}$$

and can safely be used for embedding optimization.

Estimation of history dependence by fitting a generalized linear model (GLM). Another approach to the estimation history dependence is to model the dependence of neural spiking onto past spikes explicitly, and to fit model parameters to maximize the likelihood of the observed spiking activity [13]. For a given probability distribution $p(x_t|\boldsymbol{x}_t^{-T},\nu)$ of the model with parameters parameters ν, the conditional entropy can be estimated as

$$\hat{H}(X|\boldsymbol{X}^{-T},\nu) = \frac{1}{N}\sum_{n=1}^{N}\log_2 p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T},\nu)^{-1} \tag{52}$$

which one can plug into Eq (1) to obtain an estimate of the history dependence. The strong law of large numbers [48] ensures that if the model is correct, i.e. $p(x_t|\boldsymbol{x}_t^{-T},\nu)=p(x_t|\boldsymbol{x}_t^{-T})$ for all t, this estimator converges to the entropy $H(X|\boldsymbol{X}^{-T})$ for $N\to\infty$. However, any deviations from the true distribution due to an incorrect model will lead to an underestimation of history dependence, similar to choosing an insufficient embedding. Therefore, model parameters should be chosen to maximize the history dependence, or to maximize the likelihood

$$\nu^* = \arg\max_\nu \sum_{n=1}^{N}\log_2 p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T},\nu). \tag{53}$$

We here consider a generalized linear model (GLM) with exponential link function that has successfully been applied to make predictions in neural spiking data [12] and can be used for the estimation of directed, causal information [13]. In a GLM with past dependencies, the spiking probability at time t is described by the instantaneous rate or conditional intensity function

$$\lambda(t|\boldsymbol{x}_t^{-T},\nu) = \lim_{\delta t\to 0}\frac{p(\hat{t}\in[t,t+\delta t]|\boldsymbol{x}_t^{-T},\nu)}{\delta t}. \tag{54}$$

Since we discretize spiking activity in time as spiking or non-spiking in a small time window Δt, the spiking probability is given by the binomial probability

$$p(x_t{=}1|\boldsymbol{x}_t^{-T},\nu) = \frac{\lambda(t|\boldsymbol{x}_t^{-T},\nu)\Delta t}{1+\lambda(t|\boldsymbol{x}_t^{-T},\nu)\Delta t}. \tag{55}$$

The idea of the GLM is that past events contribute independently to the probability of spiking, such that the conditional intensity function factorizes over their contributions. Hence, it can be written as

$$\lambda(t|\boldsymbol{x}_t^{-T},\mu,\boldsymbol{h}) = \exp\left(\mu + \sum_{j=1}^{d} h_j x_{t,j}^{-T}\right), \tag{56}$$

where h_j gives the contribution of past activity $x_{t,j}^{-T}$ in past time bin j to the firing rate, and μ is an offset that is adapted to match the average firing rate.
Although fitting GLM parameters is more data-efficient than computing non-parametric estimates, overfitting may occur for limited data and high embedding dimensions d, such that d cannot be chosen arbitrarily high. In order to estimate a maximum of history dependence for limited d, we apply the same type of binary past embedding as we use for the other estimators, and optimize the embedding parameters by minimizing the Bayesian information criterion [54]. In particular, for given past range T, we choose embedding parameters d^*, κ^* that minimize

$$\mathrm{BIC}(d,\kappa) = (d+1)\log_2 N - 2\mathcal{L}^*(d,\kappa), \tag{57}$$

where N is the number of samples and

$$\mathcal{L}^*(d,\kappa) = \sum_{n=1}^{N}\log_2 p(x_{t_n}|\boldsymbol{x}_{t_n,d,\kappa}^{-T},\mu^*,\boldsymbol{h}^*) \tag{58}$$

is the maximized log-likelihood of the recorded spike sequences $(x_{t_n},\boldsymbol{x}_{t_n,d,\kappa}^{-T})_{n=1}^{N}$ for optimal model parameters $\mu^*,\boldsymbol{h}^*$. We then use the optimized embedding parameters to estimate the conditional entropy according to

$$\hat{H}_{\mathrm{GLM}}(X|\boldsymbol{X}_{d^*,\kappa^*}^{-T}) = -\frac{1}{N}\mathcal{L}^*(d^*,\kappa^*), \tag{59}$$

which results in the GLM estimator of history dependence

$$\hat{R}_{\mathrm{GLM}}(T) = 1 - \frac{\hat{H}_{\mathrm{GLM}}(X|\boldsymbol{X}_{d^*,\kappa^*}^{-T})}{\hat{H}(X)}. \tag{60}$$

Bootstrap confidence intervals. In order to estimate confidence intervals of estimates $\hat{R}(T,\theta)$ for given past embeddings, we apply the *blocks of blocks* bootstrapping method [55]. To obtain bootstrap samples, we first compute all the binary sequences $(x_{t_n},\boldsymbol{x}_{t_n,\theta}^{-T})$ for $n=1,...,N$ that result from discretizing the spike recording in N time steps Δt and applying the past embedding. We then randomly draw N/l blocks of length l of the recorded binary sequences such that the total number of redrawn sequences is the same as the in the original data. We choose l to be the average inter-spike-interval (ISI) in units of time steps Δt, i.e. $l = 1/(r\Delta t)$ with average firing rate r. Sampling

successive sequences over the typical ISI ensures that bootstrapping samples are representative of the original data, while also providing a high number of distinct blocks that can be drawn.
The different estimators (but not the bias criterion) are then applied to each bootstrapping sample to obtain confidence intervals of the estimators. Instead of computing the 95% confidence interval via the 2.5 and 97.5 percentiles of the bootstrapped estimates, we assumed a Gaussian distribution and approximated the interval via $[\hat{R}(T,\theta) - 2\hat{\sigma}_R(T,\theta), \hat{R}(T,\theta) + 2\hat{\sigma}_R(T,\theta)]$, where $\hat{\sigma}_R(T,\theta)$ is the standard deviation over the bootstrapped estimates.
We found that the true standard deviation of estimates for the model neuron was well estimated by the bootstrapping procedure, irrespective of the recording length (supplementary Fig S10). Furthermore, we simulated 100 recordings of the same recording length, and for each computed confidence interval for the maximum history dependence $R_{\max}$ of Eq (18). By measuring how often the model's true value for the same embedding was included in these intervals, we found that the Gaussian confidence intervals are indeed close to the claimed confidence level (supplementary Fig S10). This indicates that the bootstrap confidence intervals approximate well the uncertainty associated with estimates of history dependence.

Cross-validation. For small recording lengths, embedding optimization may cause overfitting through the maximization of variable estimates (supplementary Fig S1). To avoid this type of overestimation, we apply cross-validation, i.e. we optimize embeddings over the first half of the recording, and evaluate estimates for the optimal past embedding on the second half. We chose this separation of training and evaluation data sets, because it allows the fastest computation of binary sequences $(x_{t_n}, \boldsymbol{x}_{t_n,\theta}^{-T})$ for the different embeddings during optimization. We found that none of the cross-validated embedding-optimized estimates were systematically overestimating the true history dependence for the benchmark model for recordings as short as three minutes (supplementary Fig S1). Therefore, cross-validation allows to apply embedding optimization to estimate history dependence even for very short recordings.

Benchmark neuron model

Generalized leaky integrate-and-fire neuron with spike-frequency adaptation. As a benchmark model, we chose a generalized leaky integrate-and-fire model (GLIF) with an additional adaptation filter ξ (GLIF-ξ) that captures spike-frequency adaptation over 20 seconds [31].
For a standard leaky integrate-and-fire neuron, the neuron's membrane is formalized as an RC circuit, where the cell's lipid membrane is modeled as a capacitance C, and the ion channels as a resistance that admits a leak current with effective conductance g_L. Hence, the temporal evolution of the membrane's voltage V is governed by

$$C\dot{V} = -g_L(V - V_R) + I_{\text{ext}}(t). \tag{61}$$

Here, V_R denotes the resting potential and $I_{\text{ext}}(t)$ external currents that are induced by some external drive. The neuron emits an action potential (spike) once the neuron crosses a voltage threshold V_T, where a spike is described as a delta pulse at the time of emission $\hat{t}$. After spike emission, the neuron returns to a reset potential V_0. Here, we do not incorporate an explicit refractory period, because inter-spike-intervals in the simulation were all larger than 10ms. For constant input current I_{ext}, integrating Eq (61) yields the membrane potential between two spiking events

$$V(t) = V_\infty + (V_0 - V_\infty)e^{-\gamma(t-\hat{t}_0)}, \tag{62}$$

where $\hat{t}_0$ is the time of the most recent spike, $\gamma = g_L/C$ the inverse membrane timescale and $V_\infty = V_R + I_{\text{ext}}/\gamma$ the equilibrium potential.
In contrast to the LIF, the GLIF models the spike emission with a soft spiking threshold. To do that, spiking is described by an inhomogeneous Poisson process, where the spiking probability in a time

window of width $\delta t \ll 1$ is given by

$$p(\hat{t} \in [t, t+\delta t]) = 1 - \exp\left(\int_t^{t+\delta t} \lambda(s) ds\right) \approx \lambda(t)\delta t. \tag{63}$$

Here, the spiking probability is governed by the time dependent firing rate

$$\lambda(t) = \lambda_0 \exp\left(\frac{V(t) - V_T(t)}{\Delta V}\right). \tag{64}$$

The idea is that once the membrane potential $V(t)$ approaches the firing threshold $V_T(t)$, the firing probability increases exponentially, where the exponential increase is modulated by $1/\Delta V$. For $\Delta V \to 0$, we recover the deterministic LIF, while for larger ΔV the emission becomes increasingly random.
In the GLIF-ξ, the otherwise constant threshold V_T^* is modulated by the neuron's own past activity according to

$$V_T(t) = V_T^* + \sum_{\hat{t}_j < t} \xi(t - \hat{t}_j). \tag{65}$$

Thus, depending on their spike times $\hat{t}_j$, emitted action potentials increase or decrease the threshold additively and independently according to an adaptation filter $\xi(t)$. Thereby $\xi(t) = 0$ for $t < 0$ to consider effects of action potentials that were emitted in the past only. In the experiments conducted in [31], the following functional form for the adaptation filter was extracted:

$$\xi(s) = \begin{cases} a_\xi & \text{, if } 0 < s \leq T_\xi \\ a_\xi \left(\frac{s}{T_\xi}\right)^{-\beta_\xi} & \text{, if } T_\xi < s < 22\,\text{s}. \end{cases} \tag{66}$$

The filter is an effective model not only for the measured increase in firing threshold, but also for spike-triggered currents that reduce the membrane potential. When mapped to the effective adaptation filter ξ, it turned out that past spikes lead to a decrease in firing probability that is approximately constant over a period $T_\xi = 8.3\,\text{ms}$, after which it decays like a power-law with exponent $\beta_\xi = 0.93$, until the contributions are set to zero after 22 s.

Simulation details. In order to ensure stationarity, we simulated the model neuron exposed to a constant external current $I_{\text{ext}} = const.$ over a total duration of $T_{\text{rec}} = 900\,\text{min}$. Thereby, the current I_{ext} was chosen such that the neuron fired with a realistic average firing rate of 4 Hz. During the simulation, Eq (61) was integrated using simple Runge-Kutta integration with an integration time step of $\delta t = 0.5$ ms. At every time step, random spiking was modeled as a binary variable with probability as in Eq (63). After a burning-in time of 100 s, spike times were recorded and used for the estimation of history dependence. The detailed simulation parameters can be found in Table 1.

Computation of the total history dependence. In order to determine the total history dependence in the simulated spiking activity, we computed the conditional entropy $H(X|\boldsymbol{X}^{-\infty})$ from the conditional spiking probability in Eq (63) that was used for the simulation. Note that this is only possible because of the constant input current, otherwise the conditional spiking probability would also capture information about the external input.
Since the conditional probability of spiking used in the simulation computes the probability in a simulation step $\delta t = 0.5\,\text{ms}$, we first have to transform this to a probability of spiking in the analysis time step $\Delta t = 5\,\text{ms}$. To do so, we compute the probability of no spike in a time step $[t, t+\Delta t]$ according to

$$p_{\text{sim}}(x_t{=}0|\boldsymbol{x}_t^{-\infty}) = \prod_{j=1}^{\Delta t/\delta t} [1 - \tilde{\lambda}(t + (j-1)\delta t)\delta t], \tag{67}$$

Term	Description	Value	Units
λ_0	Latency	2.0	ms^{-1}
$1/\gamma$	Membrane timescale	15.3	ms
V_∞	Equilibrium potential	-45.9	mV
V_0	Reset potential	-38.8	mV
V_T^*	Firing threshold baseline	-51.9	mV
ΔV	Firing threshold sharpness	0.75	mV
α_ξ	Magnitude of the effective adaptation filter ξ	19.3	mV
β_ξ	Scaling exponent of the effective adaptation filter ξ	0.93	-
T_ξ	Cutoff of the effective adaptation filter ξ	8.3	ms
δt	Simulation step	0.5	ms

Table 1. Simulation parameters of the GLIF-ξ model, originally extracted from experimental recordings of (n=14) L5 pyramidal neurons [31].

and then compute the probability of at least one spike by $p(x_t{=}1|\boldsymbol{x}_t^{-\infty}) = 1 - p(x_t{=}0|\boldsymbol{x}_t^{-\infty})$. Here, the rate $\tilde{\lambda}(t)$ is computed as $\lambda(t)$ in Eq (64), but only with respect to past spikes that are emitted at times $\hat{t} < t$. This is because no spike that occurs within $[t, t+\Delta t]$ must be considered when computing $p_{\mathrm{sim}}(x_t{=}0|\boldsymbol{x}_t^{-\infty})$.
For sufficiently long simulations, one can make use of the SLLN to compute the conditional entropy

$$H_{\mathrm{sim}}(X|\boldsymbol{X}^{-\infty}) = -\frac{1}{N}\sum_{n=1}^{N} \log_2 p_{\mathrm{sim}}(x_{t_n}|\boldsymbol{x}_{t_n}^{-\infty}), \tag{68}$$

and thus the total history dependence

$$R_{\mathrm{tot}} = 1 - \frac{H_{\mathrm{sim}}(X|\boldsymbol{X}^{-\infty})}{\hat{H}(X)}, \tag{69}$$

which gives an upper bound to the history dependence for any past embedding.

Computation of history dependence for given past embedding. To compute history dependence for given past embedding, we use that the model neuron can be well approximated by a generalized linear model (GLM) within the parameter regime of our simulation. We can then fit a GLM to the simulated data for the given past embedding T, d, κ to obtain a good approximation of the corresponding true history dependence $R(T, d, \kappa)$.
To map the model neuron to a GLM, we plug the membrane and threshold dynamics of Eq (62) and Eq (65) into the equation for the firing rate Eq(64), i.e.

$$\lambda(t) = \exp\left(\log\lambda_0 + V_\infty - V_T^* + \sum_{\hat{t}_j<t} \xi(t-\hat{t}_j) + (V_0 - V_\infty)e^{-\gamma(t-\hat{t}_0)}\right). \tag{70}$$

For the parameters used in the simulation, the decay time of the reset term $V_0 - V_\infty$ is $1/\gamma = 15.3\,\mathrm{ms}$. When compared to the minimum and mean inter-spike intervals of $\mathrm{ISI}_{\mathrm{min}} = 25,\mathrm{ms}$ and $\overline{\mathrm{ISI}} = 248\,\mathrm{ms}$, it is apparent that the probability for two spikes to occur within the decay time window is negligibly small. Therefore, one can safely approximate

$$(V_0 - V_\infty)e^{-\gamma(t-\hat{t}_0)} \approx \sum_{\hat{t}_j<t} (V_0 - V_\infty)e^{-\gamma(t-\hat{t}_j)}, \tag{71}$$

i.e. describing the potential reset after a spike as independent of other past spikes, because contributions beyond the last spike ($j > 0$) are effectively zero. Using the above approximation, one can formulate

the rate as in a generalized linear model with

$$\lambda(t) = \exp\left(\mu \sum_{j=1}^{d} h_j x_{t,j}^{-}\right), \tag{72}$$

where

$$\mu = \log \lambda_0 + V_\infty - V_T^* \tag{73}$$

$$h_j = \xi(j\delta t) + (V_0 - V_\infty)e^{-\gamma j \delta t}, \tag{74}$$

and $x_{t,j}^{-} \in \{0, 1\}$ indicates whether the neuron spiked in $[t - j\delta t, t - (j+1)\delta t]$. Therefore, the true spiking probability of the model is well described by a GLM.
We use this relation to approximate the history dependence $R(T, d, \kappa)$ for any past embedding T, d, κ with a GLM with the same past embedding. Since in that case the parameters μ and $\boldsymbol{h}$ are not known, we fitted them to the simulated 900 minute recording via maximum likelihood (see above) and computed the history dependence according to

$$\hat{R}_{\mathrm{GLM}}(T, d, \kappa) = 1 - \frac{\hat{H}_{\mathrm{GLM}}(X|\boldsymbol{X}_{d,\kappa}^{-T})}{\hat{H}(X)}. \tag{75}$$

Computation of history dependence as a function of the past range. To approximate the model's true history dependence $R(T)$, for each T we computed GLM estimates $\hat{R}_{\mathrm{GLM}}(T, d, \kappa)$ (Eq 75) for a varying number of past bins $d \in [25, 50, 75, 100, 125, 150]$. For each d, the scaling κ was chosen such that the size of the first past bin was equal or less than 0.5 ms. To save computation time, and to reduce the effect of overfitting, the GLM parameters where fitted on 300 minutes of the simulation, whereas estimates $\hat{R}_{\mathrm{GLM}}(T, d, \kappa)$ were computed on the full 900 minutes of the simulated recording. For each T, we then chose the highest estimate $\hat{R}_{\mathrm{GLM}}(T, d, \kappa)$ among the estimates for different d as the best estimate of the true $R(T)$.

Experimental recordings

We analyzed neural spike trains from *in vitro* recordings of rat cortical cultures and salamander retina, as well as *in vivo* recordings in rat dorsal hippocampus (layer CA1) and mouse primary visual cortex. Data from salamander retina were recorded in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health, and the protocol was approved by the Institutional Animal Care and Use Committee (IACUC) of Princeton University (Protocol Number: 1828). The rat dorsal hippocampus experimental protocols were approved by the Institutional Animal Care and Use Committee of Rutgers University [32, 33]. Data from mouse primary visual cortex were recorded according to the UK Animals Scientific Procedures Act (1986).

For all recordings, we only analyzed neurons with firing rates between 0.5 Hz and 10 Hz to exclude the extremes of either inactive neurons or neurons with very high firing rate.

Rat cortical culture. Neurons were extracted from rat cortex (1st day postpartum) and recorded *in vitro* on an electrode array 2-3 weeks after plating day. We took data from five consecutive sessions (`L_Prg035_txt_nounstim.txt`, `L_Prg036_txt_nounstim.txt`, ..., `L_Prg039_txt_nounstim.txt`) with a total duration of about $T_{\mathrm{rec}} \approx 203$ min. However, we only analyzed the first 90 minutes to make the results comparable to the other recorded systems. We analyzed in total $n = 48$ neurons that satisfied our requirement on the firing rate. More details on the recording procedure can be found in [56], and details on the data set proper can be found in [36].

Symbol	Value	Settings variable name	Description
Δt	0.005	`embedding_step_size`	Time step (in seconds) for the discretization of neural spiking activity.
d	$1, 2, ..., d_{\max}$	`embedding_number_of_bins_set`	Set of embedding dimensions.
N_κ	10	`number_of_scalings`	Number of linearly spaced values of the exponential scaling κ.
$\tau_{1,\min}$	0.005	`min_first_bin_size`	Minimum bin size (in seconds) of the first past bin.
$\Delta\kappa_{\min}$	0.01	`min_step_for_scaling`	Minimum required difference between two values of κ.
p	0.05	`bbc_tolerance`	Tolerance for the acceptance of estimates for BBC.
-	`False`	`cross_validated_optimization`	Is cross-validation used for optimization or not.
-	250	`number_of_bootstraps_R_max`	Number of bootstrap samples used to estimate $\sigma_{\hat{R}_{\max}}$.
l	$1/r\Delta t$	`block_length_l`	Block length used for blocks-of-blocks bootstrapping.
-	`all`	`estimation_method`	Estimators for which embeddings are optimized (BBC, Shuffling)

Table 2. Parameters used for embedding optimization. To facilitate reproduction, we added the settings variable names of the parameters as they are used in the toolbox [29].

Salamander retina. Spikes from larval tiger salamander retinal ganglion cells were recorded *in vitro* by extracting the entire retina on an electrode array [57], while a non-repeated natural movie (leaves moving in the wind) was projected onto the retina. The recording had a total length of about $T_{\text{rec}} \approx 82\,\text{min}$, and we analyzed in total $n = 111$ neurons that satisfied our requirement on the firing rate. More details on the recording procedure and the data set can be found in [34, 35]. The spike recording as obtained from the Dryad database [34].

Rat dorsal hippocampus (layer CA1). We evaluated spike trains from a multichannel simultaneous recording made from layer CA1 of the right dorsal hippocampus of a Long-Evans rat during an open field task (data set ec014.277). The data-set provided sorted spikes from 8 shanks with 64 channels. The recording had a total length of about $T_{\text{rec}} \approx 90\,\text{min}$. We analyzed in total $n = 28$ neurons that were indicated as single units and satisfied our requirement on the firing rate. More details on the experimental procedure and the data set can be found in [32, 33]. The spike recording was obtained from the NSF-founded CRCNS data sharing website.

Mouse primary visual cortex. Neurons were recorded *in vivo* during spontaneous behavior, while face expressions were monitored. Recordings were obtained by 8 simultaneously implanted Neuropixel probes, and neurons were located using the location of the electrode contacts provided in [37], and the Allen Mouse Common Coordinate Framework [58]. We analyzed in total $n = 142$ neurons from the rat "Waksman" that belonged to primary visual cortex (irrespective of their layer) and satisfied our requirement on the firing rate. Second, we only selected neurons that were recorded for more than $T_{\text{rec}} \approx 40\,\text{min}$ (difference between the last and first recorded spike time). Details on the recording procedure and the data set can be found in [59] and [37].

Parameters used for embedding optimization

The embedding dimension or number of bins was varied in a range $d \in [1, d_{\max}]$, where $d_{\max}$ was either $d_{\max} = 20$, $d_{\max} = 5$ (max five bins) or $d_{\max} = 1$ (one bin). During embedding optimization, we

explored $N_\kappa = 10$ linearly spaced values of the exponential scaling κ within a range $[0, \kappa_{\max}(d)]$. The maximum $\kappa_{\max}(d)$ was chosen for each number of bins $d \in [1, d_{\max}]$ such that the bin size of the first past bin was equal to a minimum bin size, i.e. $\tau_1 = \tau_{1,\min}$, which we chose to be equal to the time step $\tau_{1,\min} = \Delta t = 5\,\text{ms}$. To save computation time, we did not consider any embeddings with $\kappa > 0$ if the past range T and d were such that $\tau_1(\kappa_{\max}(d)) \leq \Delta t$ for $\kappa = 0$. Similarly, for given T and each d, we neglected values of κ during embedding optimization if the difference $\Delta\kappa$ to the previous value of κ was less than $\Delta\kappa_{\min} = 0.01$.

Details to Figure 3A and 3B. In Figure 3A and 3B, we applied the ML, NSB, BBC and Shuffling estimators for $R(d)$ to a simulated recording of 90 minutes. Embedding parameters were $T = d \cdot \tau$ and $\kappa = 0$, with $\tau = 20\,\text{ms}$ and $d \in [1, 100]$. Since the goal was to show the properties of the estimators, confidence intervals were estimated from 50 repeated 90 minute simulations instead of bootstrapping samples from the same recording. Each simulation had a burning in period of 100 seconds. To estimate the true $R(d)$, the GLM was fitted and evaluated on a 900 minute recording.

Details to Figure 3C. In Figure 3C, history dependence $R(T)$ was estimated on a 90 minute recording for 57 different values of T in a range $T \in [10\,\text{ms}, 3\,\text{s}]$. Embedding-optimized estimates were computed with $d_{\max} = 20$, and 95% confidence intervals were computed using the standard deviation over $n = 100$ blocks-of-blocks bootstrapping samples (see Bootstrap confidence intervals). To estimate the true $R(T, d^*, \kappa^*)$ for the optimized embedding parameters d^*, κ^* with either BBC or Shuffling, a GLM was fitted for the same embedding parameters on a 300 minute recording and evaluated on 900 minutes recording for the estimation of R. See above on how we computed the best estimate of $R(T)$.

Details to Figure 4. For Figure 4, history dependence $R(T)$ was estimated for 61 different values of T in a range $T \in [10\,\text{ms}, 5\,\text{s}]$. For each recording, we only analyzed the first 90 minutes to have a comparable recording length. For embedding optimization, we used $d_{\max} = 20$ as a default for BBC and Shuffling, and compared the estimates with the Shuffling estimator optimized for $d_{\max} = 5$ (max five bins) and $d_{\max} = 1$ (one bin). For the GLM, we only estimated $R(\hat{T}_D)$ for the temporal depth $\hat{T}_D$ that was estimated with BBC. To optimize the estimate, we computed GLM estimates $\hat{R}(\hat{T}_D)$ with the optimal embedding found by BBC, and for varying embedding dimension $d \in [1, 2, 3, .., 20, 25, 30, 35, 40, 45, 50]$, where for each d we chose κ such that $\tau_1 = \Delta t$. We then chose the embedding that minimized the BIC, and took the corresponding estimate $\hat{R}(\hat{T}_D)$ as a best estimate for R_{tot}. For Figure 3A, we plotted only spiketrains of channels that were identified as single units. For Figure 3B, 95% confidence intervals were computed using the standard deviation over $n = 100$ blocks-of-blocks bootstrapping samples. For Figure 3C, embedding-optimized estimates with uniform embedding (κ) were computed with $d_{\max} = 20$ (BBC) or $d_{\max} = 20$ (Shuffling). Medians were computed over the $n = 28$ neurons in EC.

Details to Figure 5&6. For Figure 5 and 6, history dependence was $R(T)$ was estimated for 61 different values of T in a range $T \in [10\,\text{ms}, 5\,\text{s}]$ using the Shuffling estimator with $d_{\max} = 5$. For each recording, we only analyzed the first 40 minutes to have a comparable recording length. For Figure 5, medians of $\hat{T}_D$ and $\hat{R}_{\text{tot}}$ were computed over all neurons that were analyzed, and 95% confidence intervals on the medians were obtained by bootstrapping with $n = 10000$ resamples of the median. For Figure 6, 95% confidence intervals were computed using the standard deviation over $n = 100$ blocks-of-blocks bootstrapping samples.

Practical guidelines: How to estimate history dependence from neural spike recordings

Estimating history dependence (or any complex statistical dependency) for neural data is notoriously difficult. In the following, we address the main requirements for a practical and meaningful analysis of history dependence, and provide guidelines on how to fulfill these requirements using embedding

optimization. A toolbox for Python3 is available online [29], together with default parameters that worked best with respect to the following requirements. It is important that practitioners make sure that their data fulfill the data requirements (points 4 and 5).

1) The embedding of past-spiking activity should be individually optimized to account for very different spiking statistics. It is crucial to optimize the embedding for each neuron individually, because history dependence can strongly differ for neurons from different areas or neural systems (Fig 5), or even among neurons within a single area (see examples in Fig 6). Individual optimization enables a meaningful comparison of temporal depth and history dependency R between neurons.

2) The estimation has to capture any non-linear or higher-order statistical dependencies. Embedding optimization using both, the BBC or Shuffling estimators, is based on non-parametric estimation, in which the joint probabilities of current and past spiking are directly estimated from data. Thereby, it can account for any higher-order or non-linear dependency among all bins. In contrast, the classical generalized linear model (GLM) that is commonly used to model statistical dependencies in neural spiking activity [12, 13] does not account for higher-order dependencies. We found that the GLM recovered consistently less total history dependence R_{tot} (Fig 4D). Hence, to capture single-neuron history dependence, higher-order and non-linear dependencies are important, and thus a non-parametric approach is advantageous.

3) Estimation has to be computationally feasible even for a high number of recorded neurons. Strikingly, while higher-order and non-linear dependencies are important, the estimation of history dependence does not require high temporal resolution. Optimizing up to $d_{\mathrm{max}} = 5$ past bins with variable exponential scaling κ could account for most of the total history dependence that was estimated with up to $d_{\mathrm{max}} = 20$ bins (Fig 4D). With this reduced setup, embedding optimization is feasible within reasonable computation time. Computing embedding-optimized estimates of the history dependence $R(T)$ for 61 different values of T (for 40 minute recordings, the approach used for Fig 5 and Fig 6) took around 10 minutes for the Shuffling estimator, and about 8.5 minutes for the BBC per neuron on a single computing node. Therefore, we recommend using $d_{\mathrm{max}} = 5$ past bins when computation time is a constraint. Ideally, however, one should check for a few recordings if higher choices of d_{max} lead to different results, in order to cross-validate the choice of $d_{\mathrm{max}} = 5$ for the given data set.

4) Estimates have to be reliable lower bounds, otherwise one cannot interpret the results. It is required that embedding-optimized estimates do not systematically overestimate history dependence for any given embedding. Otherwise, one cannot guarantee that *on average* estimates are lower bounds to the total history dependence, and that an increase in history dependence for higher past ranges is not simply caused by overestimation. This guarantee is an important aspect for the interpretation of the results.

For BBC, we found that embedding-optimized estimates are unbiased if the variance of estimators is sufficiently small (supplementary Fig S1). The variance was sufficiently small for recordings of 90 minutes duration. When the variance was too high (short recordings with 3–45 minutes recording length), maximizing estimates for different embedding parameters introduced very mild overestimation due to overfitting (1–3%) (supplementary Fig S1). The overfitting can, however, be avoided by cross validation, i.e. optimizing the embedding on one half of the recording and computing estimates on the other half. *Using cross-validation*, we found that embedding-optimized BBC estimates were unbiased even for recordings as short as 3 minutes (supplementary Fig S1).

For Shuffling, we also observed overfitting, but the overestimation was small compared to the inherent systematic underestimation of Shuffling estimates. Therefore, we observed no systematic overestimation by embedding-optimized Shuffling estimates on the model neuron, even for shorter recordings (3 minutes and more). Thus, for the Shuffling estimator, we advice to apply the estimator

without cross-validation as long as recordings are sufficiently long (10 minutes and more, see next point).

5) Spike recordings must be sufficiently long (at least 10 minutes), and of similar length, in order to allow for a meaningful comparison of history dependence and temporal depth among neurons. The recording length affects the estimated total history dependence $\hat{R}_{\text{tot}}$, and especially the estimated temporal depth $\hat{T}_D$. First, this is because more data allows more complex embeddings, such that more history dependence can be captured. Second, more data reduces the variance of the estimates. The variance affects the temporal depth, because only increments in history dependence are considered that are beyond statistical fluctuations. Therefore, if the variance is high, smaller temporal depth will be estimated. We found that for shorter recordings, the estimated total history dependence (thus its amount $\hat{R}_{\text{tot}}$) was roughly the same as for 90 minutes, but the estimated temporal depth $\hat{T}_D$ was much smaller (supplementary Figs S2 and S3).

To allow for a meaningful comparison of temporal depth between neurons, one thus has to ensure that recordings are sufficiently long (in our experience at least 10 minutes), otherwise differences in temporal depth are not well resolved. Below 10 minutes, we found that the estimated temporal depth $\hat{T}_D$ could be less than half of the value that was estimated for 90 minutes, and also the estimated total history dependence $\hat{R}_{\text{tot}}$ showed a notable decrease. In addition, all recordings should have comparable length to prevent that differences in history dependence or temporal depth are due to different recording lengths. Finally, it might be useful to consider additional quantities that capture the temporal aspect of history dependence. As an example, we computed the remaining history dependence $\Delta\hat{R}(T)$ after a past range of $T = 80\,\text{ms}$, which showed interesting differences between neurons in mouse primary visual cortex versus neurons in rat cortical culture and salamander retina (supplementary Fig S5). This quantity captures the amount of long-term history dependence beyond 80 ms.

Acknowledgments

We thank the the Priesemann group, especially Matthias Loidoldt, Fabian Mikulasch, Andreas Schneider, David A. Ehrlich, Jonas Dehning, F. Paul Spitzner, as well as Abdullah Makkeh, for valuable comments and for reviewing the manuscript. **Funding:** All authors received support from the Max-Planck-Society. L.R. acknowledges funding by SMARTSTART, the joint training program in computational neuroscience by the VolkswagenStiftung and the Bernstein Network. M. Wibral is employed at the Campus Institute for Dynamics of Biological Networks funded by the VolkswagenStiftung. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the PP 2205 - project number 430157073. **Author contributions:** L.R. and V.P. designed the research. L.R. conducted the research. L.R. and D.G.M. analyzed the data. L.R. and D.G.M. created the figures. All authors wrote the paper. **Competing interests:** The authors do not declare any competing interests. **Data and materials availability:** Both data and analysis code are available online. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

References

1. Barlow HB. Possible Principles Underlying the Transformations of Sensory Messages. In: Rosenblith WA, editor. Sensory Communication. The MIT Press;. p. 216–234. Available from: `http://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262518420.001.0001/upso-9780262518420-chapter-13`.

2. Press TM. Spikes — The MIT Press;. Available from: `https://mitpress.mit.edu/books/spikes`.

3. Pozzorini C, Naud R, Mensi S, Gerstner W. Temporal Whitening by Power-Law Adaptation in Neocortical Neurons;16(7):942. doi:10.1038/nn.3431.

4. Atick JJ. Could Information Theory Provide an Ecological Theory of Sensory Processing?;3(2):213–251. doi:10.1088/0954-898X$_{320}$09.

5. Lizier JT. Computation in Complex Systems. In: Lizier JT, editor. The Local Information Dynamics of Distributed Computation in Complex Systems. Springer Theses. Springer Berlin Heidelberg;. p. 13–52. Available from: `https://doi.org/10.1007/978-3-642-32952-4_2`.

6. Wibral M, Lizier JT, Vögler S, Priesemann V, Galuske R. Local Active Information Storage as a Tool to Understand Distributed Neural Information Processing;8. doi:10.3389/fninf.2014.00001.

7. Barlow H. Redundancy Reduction Revisited;12(3):241–253.

8. Murray JD, Bernacchia A, Freedman DJ, Romo R, Wallis JD, Cai X, et al. A Hierarchy of Intrinsic Timescales across Primate Cortex;17(12):1661–1663. doi:10.1038/nn.3862.

9. Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N. A Hierarchy of Temporal Receptive Windows in Human Cortex;28(10):2539–2550. doi:10.1523/JNEUROSCI.5487-07.2008.

10. Wilting J, Dehning J, Pinheiro Neto J, Rudelt L, Wibral M, Zierenberg J, et al. Operating in a Reverberating Regime Enables Rapid Tuning of Network States to Task Requirements;12. doi:10.3389/fnsys.2018.00055.

11. Archer EW, Park IM, Pillow JW. Bayesian Entropy Estimation for Binary Spike Train Data Using Parametric Prior Knowledge. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in Neural Information Processing Systems 26. Curran Associates, Inc.;. p. 1700–1708. Available from: `http://papers.nips.cc/paper/4873-bayesian-entropy-estimation-for-binary-spike-train-data-using-parametric-prior-knowle pdf`.

12. Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky EJ, et al. Spatio-Temporal Correlations and Visual Signalling in a Complete Neuronal Population;454(7207):995–999. doi:10.1038/nature07140.

13. Quinn CJ, Coleman TP, Kiyavash N, Hatsopoulos NG. Estimating the Directed Information to Infer Causal Relationships in Ensemble Neural Spike Train Recordings;30(1):17–44. doi:10.1007/s10827-010-0247-2.

14. Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W. Entropy and Information in Neural Spike Trains;80(1):197–200. doi:10.1103/PhysRevLett.80.197.

15. Panzeri S, Treves A, Schultz S, Rolls ET. On Decoding the Responses of a Population of Neurons from Short Time Windows;11(7):1553–1577. doi:10.1162/089976699300016142.

16. Brenner N, Strong SP, Koberle R, Bialek W, van Steveninck RRdR. Synergy in a Neural Code;12(7):1531–1552. doi:10.1162/089976600300015259.

17. Panzeri S, Schultz SR. A Unified Approach to the Study of Temporal, Correlational, and Rate Coding;13(6):1311–1349. doi:10.1162/08997660152002870.

18. Stetter O, Battaglia D, Soriano J, Geisel T. Model-Free Reconstruction of Excitatory Neuronal Connectivity from Calcium Imaging Signals;8(8):e1002653. doi:10.1371/journal.pcbi.1002653.

19. Panzeri S, Treves A. Analytical Estimates of Limited Sampling Biases in Different Information Measures;7(1):87–107. doi:10.1080/0954898X.1996.11978656.

20. Paninski L. Estimation of Entropy and Mutual Information;15(6):1191–1253. doi:10.1162/089976603321780272.

21. Panzeri S, Schultz SR, Treves A, Rolls ET. Correlations and the Encoding of Information in the Nervous System;266(1423):1001–1012. doi:10.1098/rspb.1999.0736.

22. Panzeri S, Senatore R, Montemurro MA, Petersen RS. Correcting for the Sampling Bias Problem in Spike Train Information Measures;98(3):1064–1072. doi:10.1152/jn.00559.2007.

23. Montemurro MA, Senatore R, Panzeri S. Tight Data-Robust Bounds to Mutual Information Combining Shuffling and Model Selection Techniques;19(11):2913–2957. doi:10.1162/neco.2007.19.11.2913.

24. Wolpert DH, Wolf DR. Estimating Functions of Probability Distributions from a Finite Set of Samples;52(6):6841–6854. doi:10.1103/PhysRevE.52.6841.

25. Nemenman I, Bialek W, de Ruyter van Steveninck R. Entropy and Information in Neural Spike Trains: Progress on the Sampling Problem;69(5):056111. doi:10.1103/PhysRevE.69.056111.

26. Archer E, Park I, Pillow J. Bayesian Entropy Estimation for Countable Discrete Distributions;15.

27. Small M. Time Series Embedding and Reconstruction. In: Applied Nonlinear Time Series Analysis: Applications in Physics, Physiology and Finance. vol. Volume 52 of World Scientific Series on Nonlinear Science Series A. WORLD SCIENTIFIC;. p. 1–46. Available from: `https://www.worldscientific.com/doi/abs/10.1142/9789812567772_0001`.

28. Palmigiano A, Geisel T, Wolf F, Battaglia D. Flexible Information Routing by Transient Synchrony;20(7):1014. doi:10.1038/nn.4569.

29. Rudelt L, Marx DG, Wibral M, Priesemann V. History Dependence Estimator;. Available from: `https://github.com/Priesemann-Group/hdestimator`.

30. Shannon CE. A Mathematical Theory of Communication;27(3):379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

31. Mensi S, Naud R, Pozzorini C, Avermann M, Petersen CCH, Gerstner W. Parameter Extraction and Classification of Three Cortical Neuron Types Reveals Two Distinct Adaptation Mechanisms;107(6):1756–1775. doi:10.1152/jn.00408.2011.

32. Mizuseki K, Sirota A, Pastalkova E, Buzsáki G. Multi-Unit Recordings from the Rat Hippocampus Made during Open Field Foraging.;. Available from: `http://crcns.org/data-sets/hc/hc-2`.

33. Mizuseki K, Sirota A, Pastalkova E, Buzsáki G. Theta Oscillations Provide Temporal Windows for Local Circuit Computation in the Entorhinal-Hippocampal Loop;64(2):267–280. doi:10.1016/j.neuron.2009.08.037.

34. Loback AR, Tkačik G, Prentice JS, Ioffe ML, J BI Michael, Marre O, et al.. Data from: Error-Robust Modes of the Retinal Population Code;. Available from: `http://datadryad.org/stash/dataset/doi:10.5061/dryad.1f1rc`.

35. Prentice JS, Marre O, Ioffe ML, Loback AR, Tkačik G, Ii MJB. Error-Robust Modes of the Retinal Population Code;12(11):e1005148. doi:10.1371/journal.pcbi.1005148.

36. Marom S. MEA Data;1. doi:10.17632/4ztc7yxngf.1.

37. Stringer C, Pachitariu M, Carandini M, Harris K. Eight-Probe Neuropixels Recordings during Spontaneous Behaviors;. Available from: `https://janelia.figshare.com/articles/dataset/Eight-probe_Neuropixels_recordings_during_spontaneous_behaviors/7739750`.

38. Brockwell PJ, Davis RA. Time Series: Theory and Methods. 2nd ed. Springer Series in Statistics. Springer-Verlag;. Available from: https://www.springer.com/de/book/9780387974293.

39. Chapeau-Blondeau F. Autocorrelation versus Entropy-Based Autoinformation for Measuring Dependence in Random Signal;380:1–18. doi:10.1016/j.physa.2007.02.077.

40. Spitzner FP, Dehning J, Wilting J, Hagemann A, Neto JP, Zierenberg J, et al. MR. Estimator, a Toolbox to Determine Intrinsic Timescales from Subsampled Spiking Activity;.

41. Wilting J, Priesemann V. Inferring Collective Dynamical States from Widely Unobserved Systems;9(1):2325. doi:10.1038/s41467-018-04725-4.

42. Wilting J, Priesemann V. Between Perfectly Critical and Fully Irregular: A Reverberating Model Captures and Predicts Cortical Spike Propagation;29(6):2759–2770. doi:10.1093/cercor/bhz049.

43. Wilting J, Priesemann V. 25 Years of Criticality in Neuroscience — Established Results, Open Controversies, Novel Concepts;58:105–111. doi:10.1016/j.conb.2019.08.002.

44. Dong DW, Atick JJ. Temporal Decorrelation: A Theory of Lagged and Nonlagged Responses in the Lateral Geniculate Nucleus;6(2):159–178. doi:10.1088/0954-898X$_{620}$03.

45. Wang XJ, Liu Y, Sanchez-Vives MV, McCormick DA. Adaptation and Temporal Decorrelation by Single Neurons in the Primary Visual Cortex;89(6):3279–3293. doi:10.1152/jn.00242.2003.

46. Moser EI, Kropff E, Moser MB. Place Cells, Grid Cells, and the Brain's Spatial Representation System;31(1):69–89. doi:10.1146/annurev.neuro.31.061307.090723.

47. Masquelier T, Deco G. Network Bursting Dynamics in Excitatory Cortical Neuron Cultures Results from the Combination of Different Adaptive Mechanism;8(10):e75824. doi:10.1371/journal.pone.0075824.

48. Meyn SP, Tweedie RL. Markov Chains and Stochastic Stability. Communications and Control Engineering. Springer-Verlag;. Available from: //www.springer.com/de/book/9781447132691.

49. Kullback S, Leibler RA. On Information and Sufficiency;22(1):79–86. doi:10.1214/aoms/1177729694.

50. Kiefer J, Wolfowitz J. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters;27(4):887–906. doi:10.1214/aoms/1177728066.

51. Antos A, Kontoyiannis I. Convergence Properties of Functional Estimates for Discrete Distributions;19(3-4):163–193. doi:10.1002/rsa.10019.

52. Treves A, Panzeri S. The Upward Bias in Measures of Information Derived from Limited Data Samples;7(2):399–407. doi:10.1162/neco.1995.7.2.399.

53. Nemenman I, Shafee F, Bialek W. Entropy and Inference, Revisited;.

54. Schwarz G. Estimating the Dimension of a Model;6(2):461–464. doi:10.1214/aos/1176344136.

55. Davison AC, Hinkley DV. Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press;. Available from: https://www.cambridge.org/core/books/bootstrap-methods-and-their-application/ED2FD043579F27952363566DC09CBD6A.

56. Shahaf G, Marom S. Learning in Networks of Cortical Neurons;21(22):8782–8788. doi:10.1523/JNEUROSCI.21-22-08782.2001.

57. Marre O, Amodei D, Deshmukh N, Sadeghi K, Soo F, Holy TE, et al. Mapping a Complete Neural Population in the Retina;32(43):14859–14873. doi:10.1523/JNEUROSCI.0723-12.2012.

58. Wang Q, Ding SL, Li Y, Royall J, Feng D, Lesnar P, et al. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas;181(4):936–953.e20. doi:10.1016/j.cell.2020.04.007.

59. Stringer C, Pachitariu M, Steinmetz N, Reddy CB, Carandini M, Harris KD. Spontaneous Behaviors Drive Multidimensional, Brainwide Activity;364(6437). doi:10.1126/science.aav7893.

Supplementary material

- Supplementary figures S1 to S10

Supplementary figures

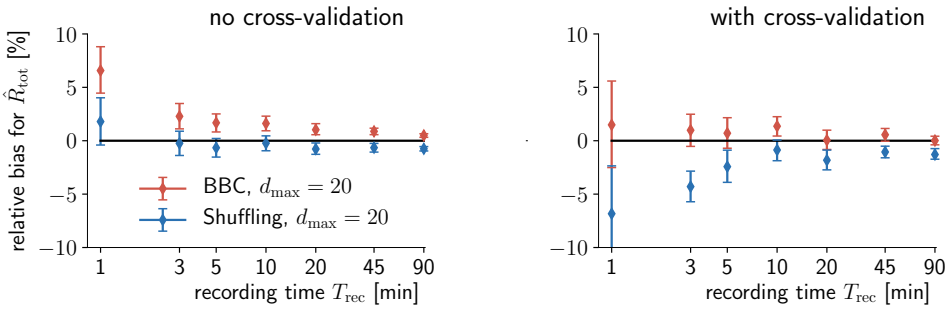

Fig S1. Embedding optimization leads to mild overfitting for short recordings, which can be avoided by cross-validation. Shown is the relative bias, i.e. the relative difference between embedding-optimized estimates $\hat{R}(T, d^*, \kappa^*)$ and the the model's true history dependence $R(T, d^*, \kappa^*)$ for the same optimized embedding parameters d^*, κ^*. The relative bias for $\hat{R}_{\mathrm{tot}}$ was computed by first averaging the relative difference $(\hat{R}(T, d^*, \kappa^*) - R(T, d^*, \kappa^*))/R(T, d^*, \kappa^*)$ for $T \in [\hat{T}_D, T_{\max}]$, and second averaging again over 30 different simulations for T_{rec} between 1 and 20 minutes, and 10 different simulations for 45 and 90 minutes. Embedding parameters were optimized for each simulation, respectively, using parameters as in Table 2. **(Left)** For BBC, the relative bias for $\hat{R}_{\mathrm{tot}}$ is zero only if recordings are sufficiently long ($\approx$ 90 minutes). When recordings are shorter, the relative bias increases, and thus estimates are mildly overestimating the model's true history dependence for the optimized embedding parameters. For Shuffling, estimates provide lower bounds to the model's true history dependence, such that the relative bias remains negative even in the presence of overfitting. **(Right)** When cross-validation is applied, i.e. embedding parameters are optimized on one half, and estimates are computed on the other half of the data, even for short recordings the bias is approximately zero for BBC, or more negative for the Shuffling estimator. Therefore, we conclude that the origin of overfitting is the selection of embedding parameters on the same data that are used for the estimation of R. Errorbars show 95 % bootstrapping confidence intervals on the mean over different simulations.

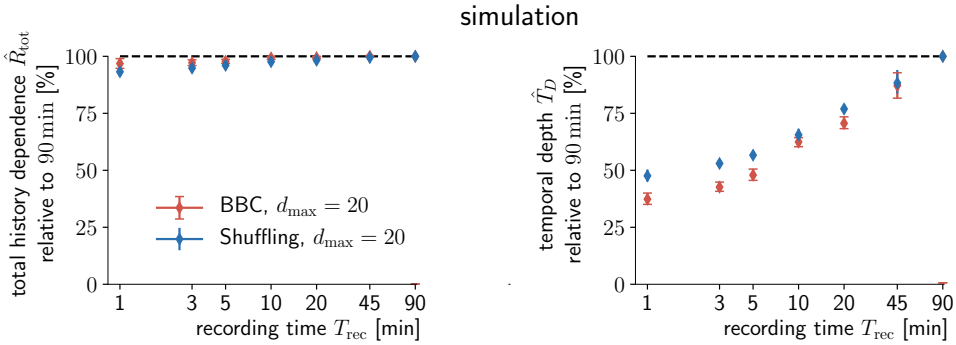

Fig S2. For the simulated neuron, recording length has little effect on estimated total history dependence, but large impact on the estimated temporal depth. (Left) Estimated total history dependence $\hat{R}_{\text{tot}}$ for different recording lengths, relative to the mean total history dependence estimated for 90 minute recordings (mean over 10 simulations). As the recording length decreases, also $\hat{R}_{\text{tot}}$ decreases. However, with only 3 minutes, one does still infer about $\approx 95\%$ of $\hat{R}_{\text{tot}}$ that one does estimate with 90 minutes of data. **(Right)** In contrast, the estimated temporal depth decreases strongly with decreasing recording length. With 3 minutes and less, only $\approx 50\%$ of the mean $\hat{T}_D$ for 90 minutes is estimated on average. Shown are mean values for 30 different simulations for T_{rec} between 1 and 20 minutes, and 10 different simulations for 45 minutes, as well as 95% confidence intervals on the mean based on bootstrapping.

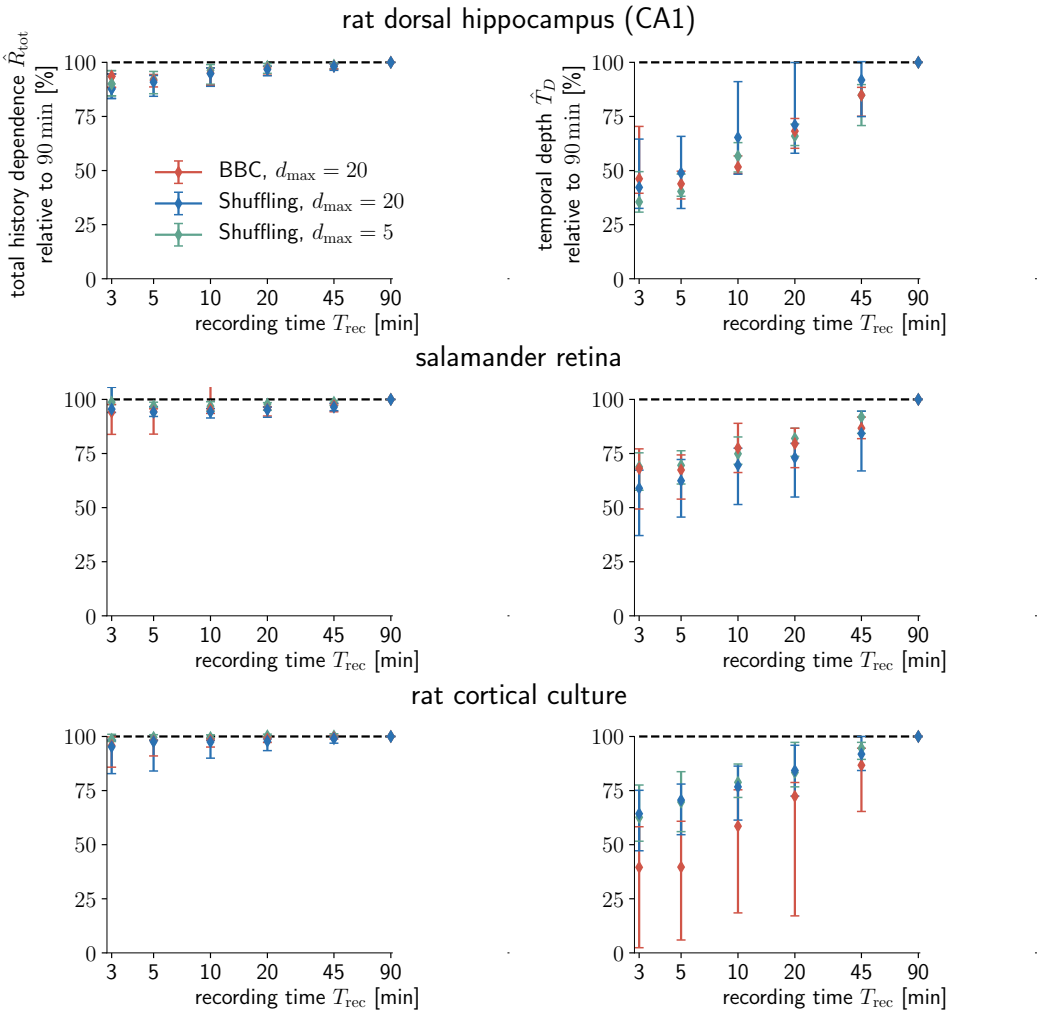

Fig S3. Also for experimental data, recording length has little effect on estimated total history dependence, but large impact on the estimated temporal depth. (Left) Estimated total history dependence $\hat{R}_{\mathrm{tot}}$ for different recording lengths, relative to the total history dependence estimated for a 90 minute recording. As long as recordings are 10 minutes or longer, one does still estimates about $\approx 95\%$ or more of $\hat{R}_{\mathrm{tot}}$ than for 90 minutes, for all three recordings. For less than 10 minutes, the estimated total history dependence decreases down to 90% (CA1), or increases again due to overfitting (retina). **(Right)** Similar to the simulated neuron, the estimated temporal depth $\hat{T}_D$ decreases strongly with decreasing recording length. With 10 minutes and more, one estimates around $\approx 50\%$ or more of $\hat{T}_D$ that is estimated on a 90 minute recording. For even shorter recordings, the median $\hat{T}_D$ drops even below $\approx 50\%$ for CA1 and rat cortical culture (BBC). Shown is the median with 95% bootstrapping confidence intervals over $n = 10$ randomly chosen neurons for each recorded system. Before taking the median over neurons, for each neuron we averaged estimates over 10 excerpts of the full recording, each with 3 or 5 minutes duration, and over 8,4 and 2 excerpts with 10, 20 and 45 minutes duration, respectively.

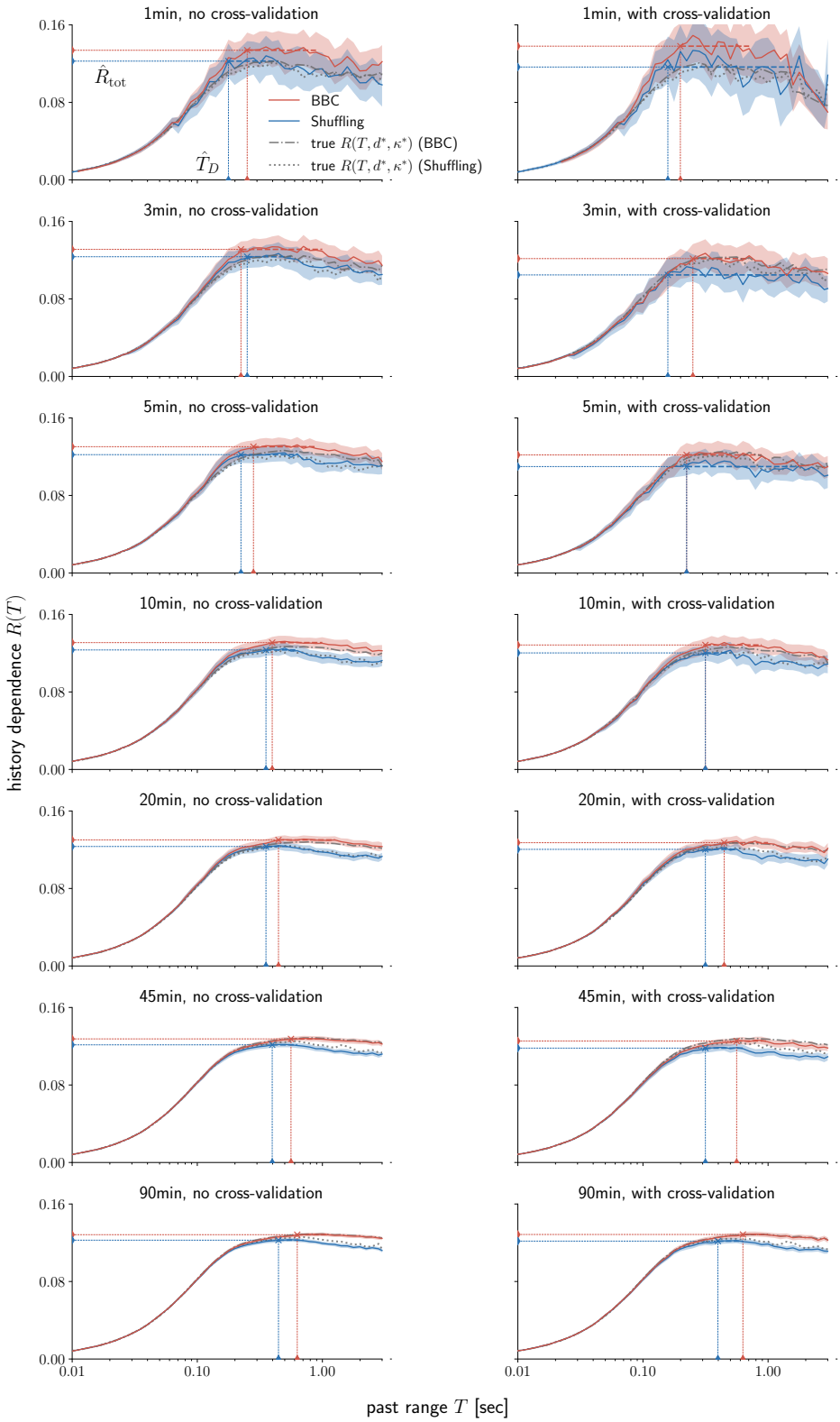

Fig S4. Estimation results for model neuron for simalated example recordings. For each neuron, we show the embedding-optimized estimates of history dependence $R(T)$ for BBC with $d_{\max} = 20$ (red) and Shuffling with $d_{\max} = 20$ (blue), as well as the ground truth for the same embeddings that were found during optimization (dashed lines). Colored dashed lines indicate the estimated temporal depth $\hat{T}_D$ and total history dependence $\hat{R}_{\text{tot}}$.

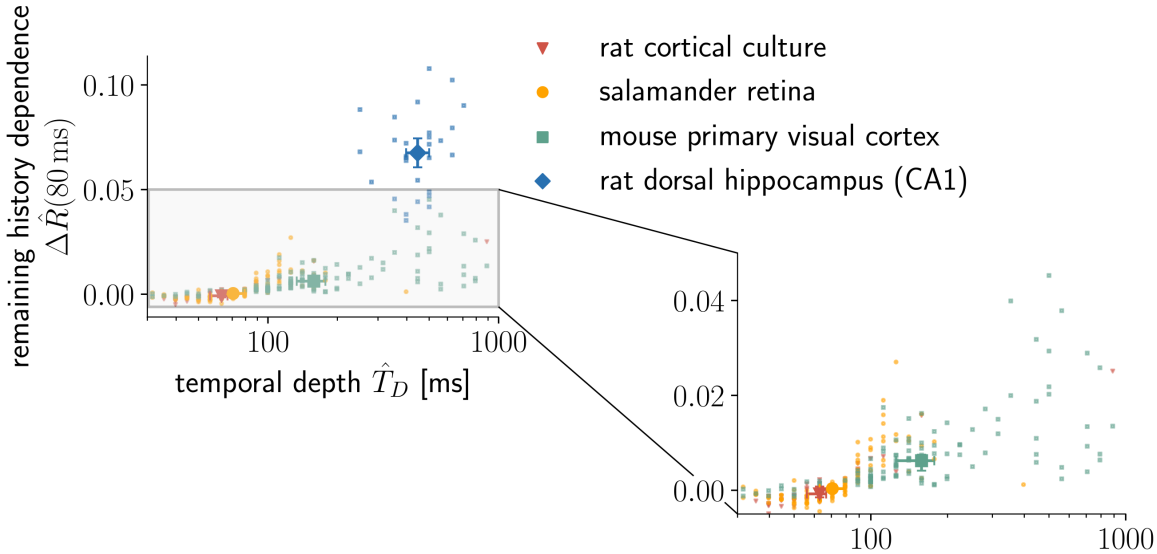

Fig S5. Remaining history dependence reveals that many neurons in mouse primary visual cortex still integrate a substantial amount of past information after 80 ms. Shown is the remaining history dependence $\Delta\hat{R}(T) \equiv \hat{R}_{\mathrm{tot}} - \hat{R}(T)$ for $T = 80\,\mathrm{ms}$ for all neurons in cortical culture, retina and primary visual cortex, as well as hippocampus layer CA1 (dots). Errorbars show median as well as 95% bootstrapping confidence intervals on the median. The $\Delta\hat{R}(T)$ is much higher for neurons in hippocampus layer CA1, but also in visual cortex it is substantially higher than in retina or cortical culture.

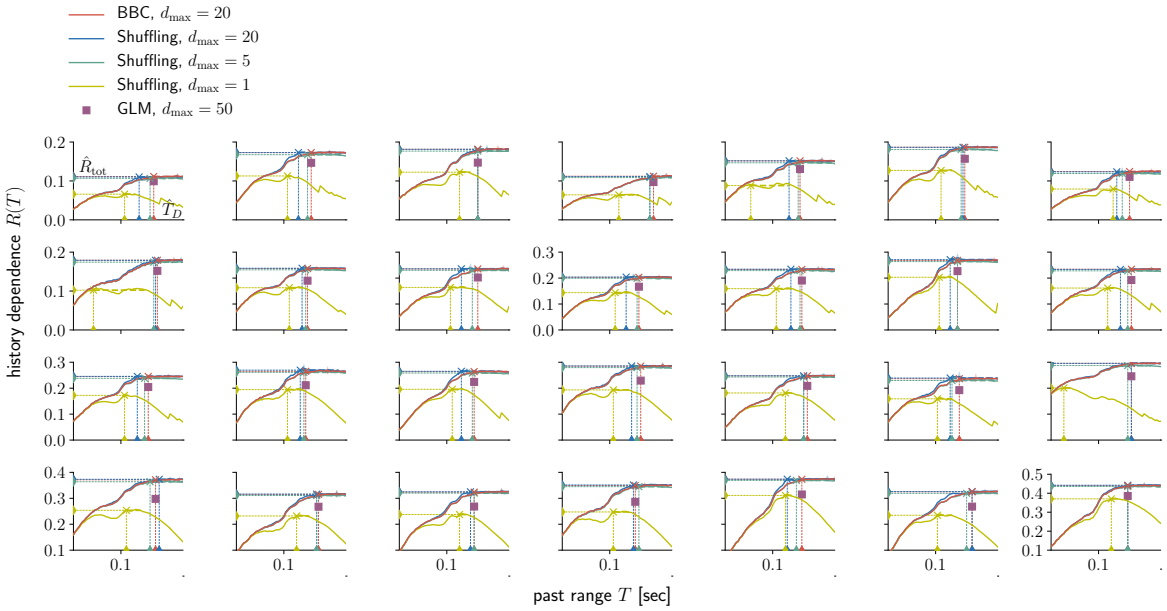

Fig S6. Estimation results for all neurons in rat dorsal hippocampus (layer CA1). For each neuron, we show the embedding-optimized estimates of history dependence $R(T)$ for BBC with $d_{\max} = 20$ (red), as well as Shuffling with $d_{\max} = 20$ (blue), $d_{\max} = 5$ (green) and $d_{\max} = 1$ (yellow). Dashed lines indicate the estimated temporal depth $\hat{T}_D$ and total history dependence $\hat{R}_{\mathrm{tot}}$. Also shown is the embedding optimized GLM estimate for the same temporal depth $\hat{T}_D$ as was found with BBC (violet square).

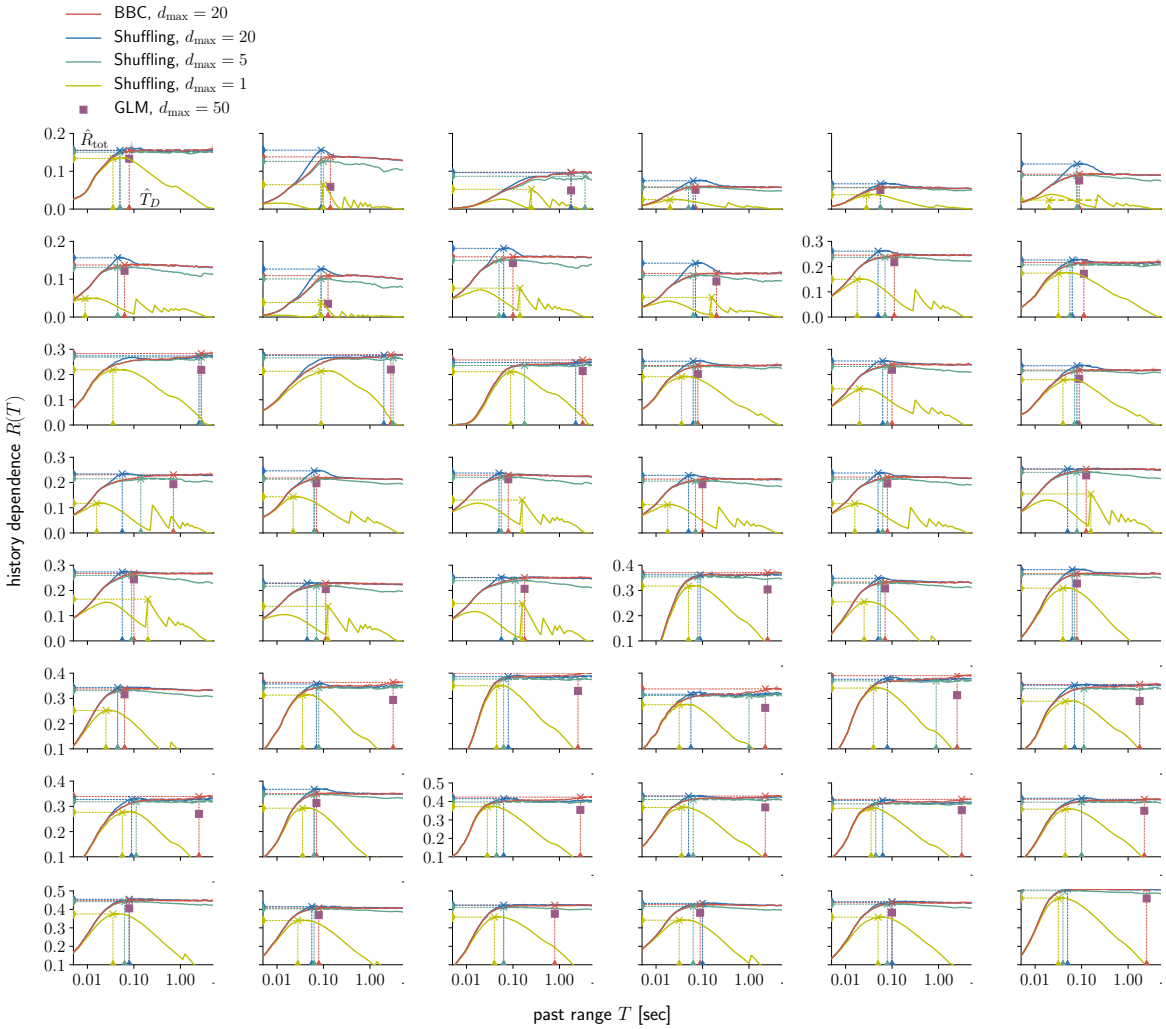

Fig S7. Estimation results for all neurons in rat cortical culture. For each neuron, we show the embedding-optimized estimates of history dependence $R(T)$ for BBC with $d_{\max} = 20$ (red), as well as Shuffling with $d_{\max} = 20$ (blue), $d_{\max} = 5$ (green) and $d_{\max} = 1$ (yellow). Dashed lines indicate the estimated temporal depth $\hat{T}_D$ and total history dependence $\hat{R}_{\text{tot}}$. Also shown is the embedding optimized GLM estimate for the same temporal depth $\hat{T}_D$ as was found with BBC (violet square).

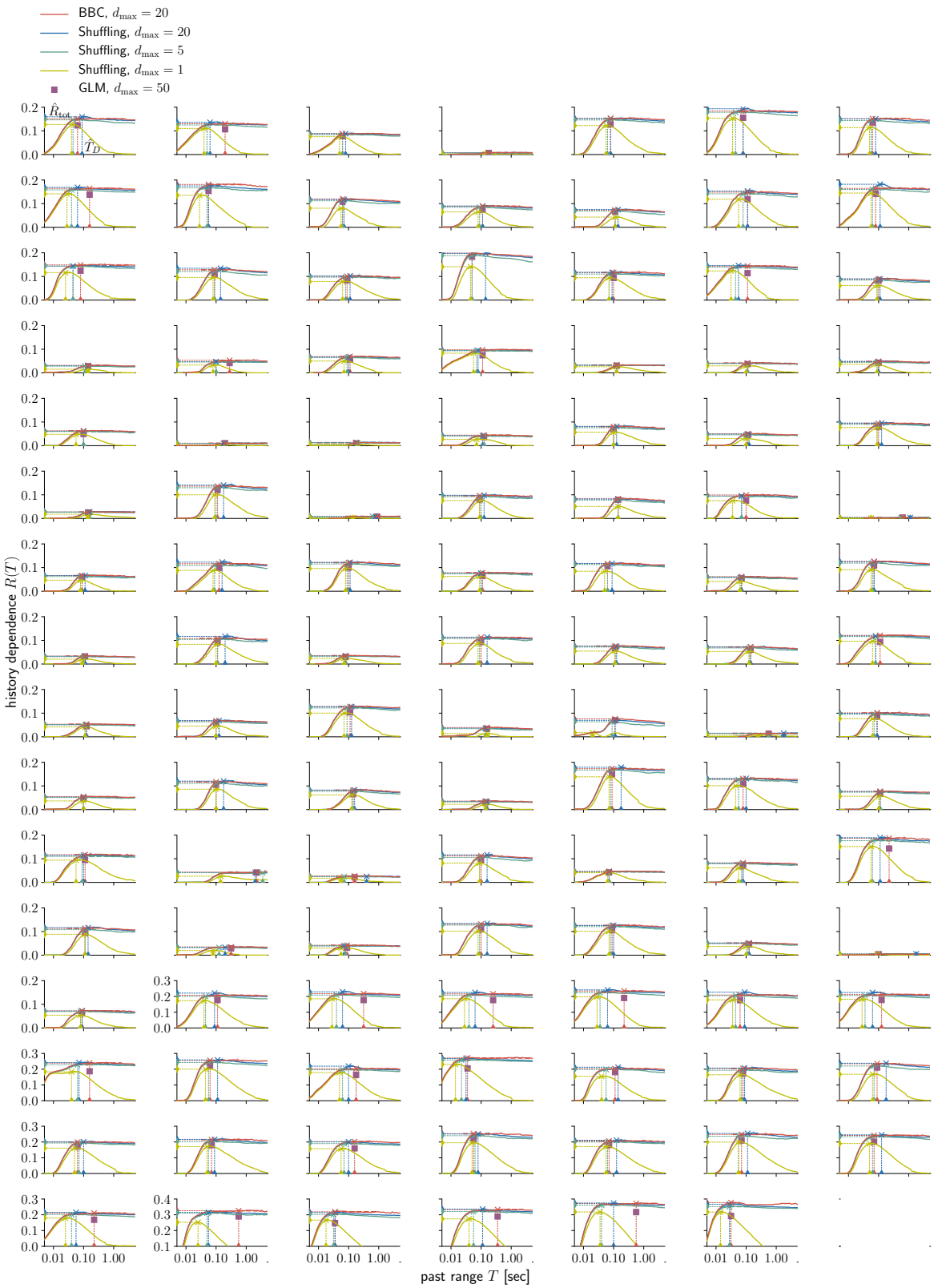

Fig S8. Estimation results for all neurons in salamander retina. For each neuron, we show the embedding-optimized estimates of history dependence $R(T)$ for BBC with $d_{\max} = 20$ (red), as well as Shuffling with $d_{\max} = 20$ (blue), $d_{\max} = 5$ (green) and $d_{\max} = 1$ (yellow). Dashed lines indicate the estimated temporal depth $\hat{T}_D$ and total history dependence $\hat{R}_{\text{tot}}$. Also shown is the embedding optimized GLM estimate for the same temporal depth $\hat{T}_D$ as was found with BBC (violet square).

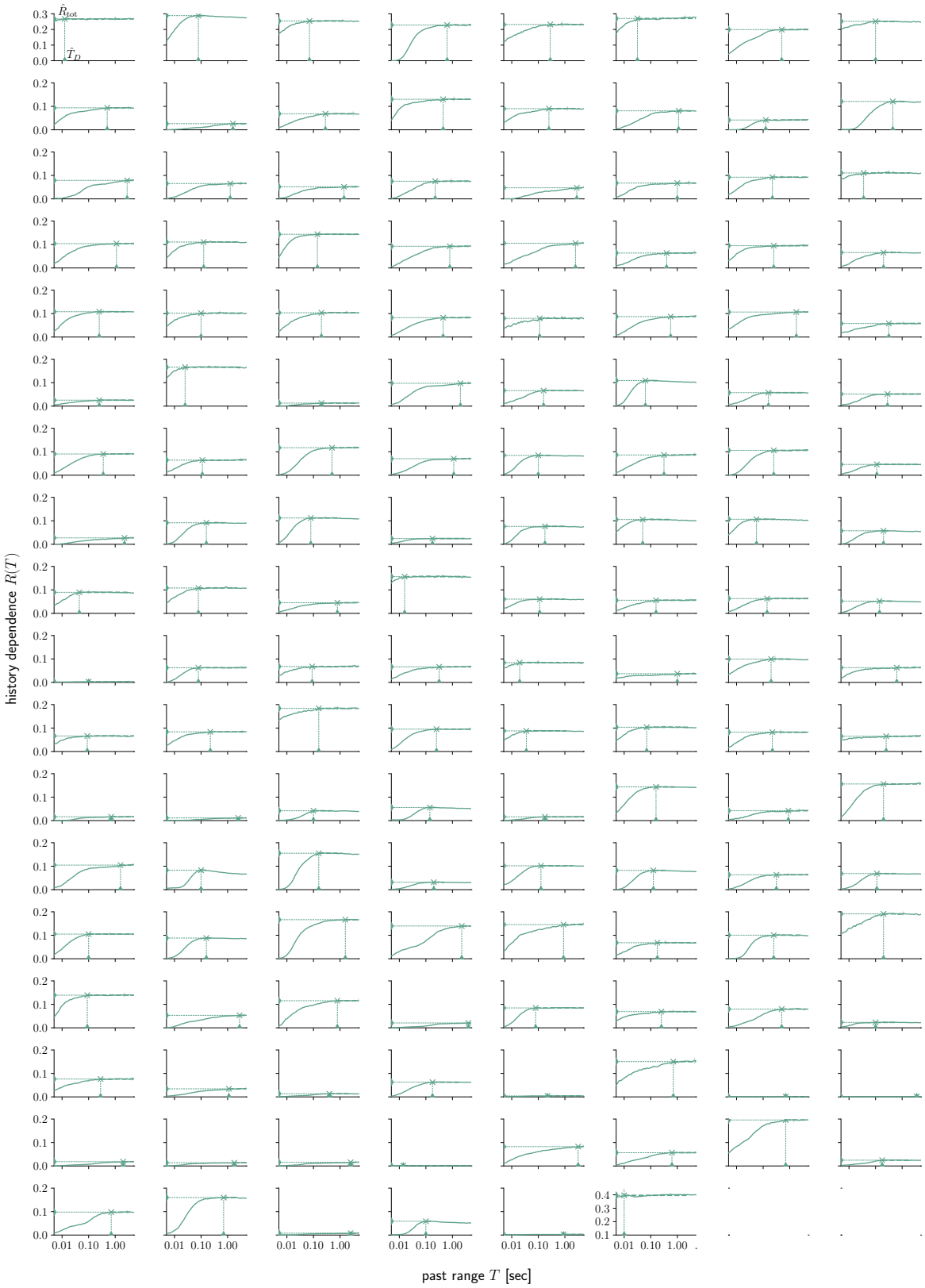

Fig S9. Estimation results for all neurons in mouse primary visual cortex (V1). For each neuron, we show the embedding-optimized Shuffling estimates of history dependence $R(T)$ for $d_{\max} = 5$. Dashed lines indicate the estimated temporal depth $\hat{T}_D$ and total history dependence $\hat{R}_{\mathrm{tot}}$.

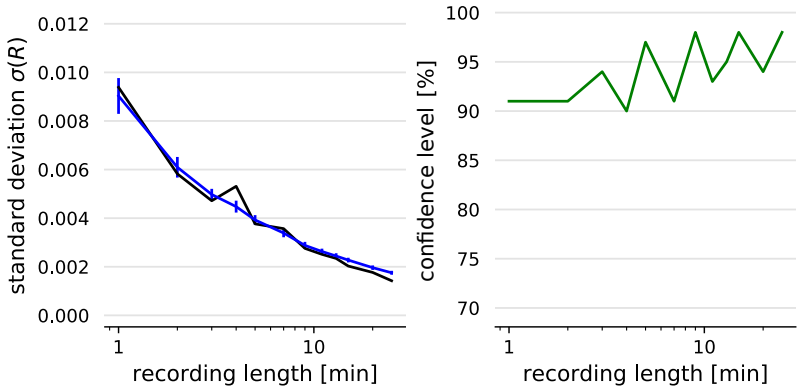

Fig S10. Bootstrapping yields accurate estimates of standard deviation and confidence intervals. (Left) The standard deviation on BBC estimates obtained from 250 blocks of blocks bootstrap samples (Materials and methods) on a single recording (blue) agrees well with the true standard deviation (black), which we estimated from 100 repeated simulations of the same recording length and embedding (black). As expected, the standard deviation decreases substantially for longer recordings. For each recording length, estimates were computed for typical optimal embedding parameters d^*, κ^* and $T = \hat{T}_D$ that were found by embedding optimization. Errorbars show mean and standard deviation of the estimated $\sigma(R)$ over the repeated simulations. **(Right)** The 95% confidence intervals based on two standard deviations $\sigma(R)$ over 250 blocks of blocks bootstrap samples have approximately the claimed confidence level. For each recording length, we computed estimates $\hat{R}$ and the bootstrapping confidence intervals on the 100 simulations, and computed the confidence level by counting how often the true value of R was contained in the confidence interval (green line). Estimates and the true value of R were computed for the same typical embedding parameters as before.