PLOS Computational Biology

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

--Manuscript Draft--

Manuscript Number:	PCOMPBIOL-D-20-01987R1
Full Title:	Embedding optimization reveals long-lasting history dependence in neural spiking activity
Short Title:	History dependence in neural spiking activity
Article Type:	Research Article
Keywords:	neural information processing, neural coding, information theory, entropy estimation, mutual information, active information storage, intrinsic timescale
Abstract:	Information processing can leave distinct footprints on the statistics of neural spiking. For example, efficient coding minimizes the statistical dependencies on the spiking history, while temporal integration of information may require the maintenance of information over different timescales. To investigate these footprints, we developed a novel approach to quantify history dependence within the spiking of a single neuron, using the mutual information between the entire past and current spiking. This measure captures how much past information is necessary to predict current spiking. In contrast, classical pairwise measures of temporal dependence like the autocorrelation capture how long—-potentially redundant—-past information can still be read out. Strikingly, we find for model neurons that our method disentangles the strength and timescale of history dependence, whereas the two are mixed in classical approaches. When applying the method to experimental data, which are necessarily of limited size, a reliable estimation of mutual information is only possible for a coarse temporal binning of past spiking, a so called past embedding. To still account for the vastly different spiking statistics and potentially long history dependence of living neurons, we developed an embedding-optimization approach that does not only vary the number and size, but also an exponential stretching of past bins. For extra-cellular spike recordings, we found that the strength and timescale of history dependence indeed can vary independently across experimental preparations. While hippocampus indicated strong and long history dependence, in visual cortex it was weak and short, while in vitro the history dependence was strong but short. This work enables an information theoretic characterization of history dependence in recorded spike trains, which captures a footprint of information processing that is beyond pairwise measures of temporal dependence. To facilitate the application of the method, we provide practical guidelines and a toolbox.
Additional Information:	
Question	**Response**
Financial Disclosure Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS Computational Biology* for specific examples. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate.	All authors received support from the Max-Planck-Society, https://www.mpg.de/de. This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the SPP 2205 Evolutionary Convergence - project number 430157073. L.R. acknowledges funding by SMARTSTART, the joint training program in computational neuroscience by the VolkswagenStiftung and the Bernstein Network, https://www.smartstart-compneuro.de/. M. W. is employed at the Campus Institute for Dynamics of Biological Networks funded by the VolkswagenStiftung, https://www.volkswagenstiftung.de. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Unfunded studies Enter: *The author(s) received no specific funding for this work.* **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? • **NO** - Include this sentence at the end of your statement: *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.* • **YES** - Specify the role(s) played. * typeset	
Competing Interests Use the instructions below to enter a competing interest statement for this submission. On behalf of all authors, disclose any competing interests that could be perceived to bias this work—acknowledging all financial support and any other relevant financial or non-financial competing interests. This statement **will appear in the published article** if the submission is accepted. Please make sure it is accurate. View published research articles from *PLOS Computational Biology* for specific examples.	The authors have declared that no competing interests exist.

NO authors have competing interests Enter: *The authors have declared that no competing interests exist.* **Authors with competing interests** Enter competing interest details beginning with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests: [insert competing interests here]* * typeset This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate and that any funding sources listed in your Funding Information later in the submission form are also declared in your Financial Disclosure statement.	
Data Availability Authors are required to make all data underlying the findings described fully available, without restriction, and from the time of publication. PLOS allows rare exceptions to address legal and ethical concerns. See the PLOS Data Policy and FAQ for detailed information. A Data Availability Statement describing where the data can be found is required at submission. Your answers to this question constitute the Data Availability Statement and **will be published in the article**, if accepted. **Important:** Stating 'data available on request from the author' is not sufficient. If your data are only available upon request, select 'No' for the first question and explain your exceptional situation in the text box.	Yes - all data are fully available without restriction

Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	
Describe where the data may be found in full sentences. If you are copying our sample text, replace any instances of XXX with the appropriate details. • If the data are **held or will be held in a public repository**, include URLs, accession numbers or DOIs. If this information will only be available after acceptance, indicate this by ticking the box below. For example: *All XXX files are available from the XXX database (accession number(s) XXX, XXX.*). • If the data are all contained **within the manuscript and/or Supporting Information files**, enter the following: *All relevant data are within the manuscript and its Supporting Information files.* • If neither of these applies but you are able to provide **details of access elsewhere**, with or without limitations, please do so. For example: *Data cannot be shared publicly because of [XXX]. Data are available from the XXX Institutional Data Access / Ethics Committee (contact via XXX) for researchers who meet the criteria for access to confidential data.* *The data underlying the results presented in the study are available from (include the name of the third party and contact information or URL).* • This text is appropriate if the data are owned by a third party and authors do not have permission to share the data. * typeset	The data underlying the results presented in the study are available online - from the CRCNS (crcns.org) data repository: http://crcns.org/data-sets/hc/hc-2 - from the Mendeley Data repository: doi:10.17632/4ztc7yxngf.1 - from the Dryad database: doi:10.5061/dryad.1f1rc - from the Janelia figshare repository: https://janelia.figshare.com/articles/dataset/Eight-probe_Neuropixels_recordings_during_spontaneous_behaviors/7739750
Additional data availability information:	

MAX-PLANCK-GESELLSCHAFT

L. Rudelt, D. G. Marx, M. Wibral, V. Priesemann

Max-Planck-Institute for Dynamics and Self-Organization
Am Faßberg 17
D-37077 Göttingen
✉ lucas.rudelt@ds.mpg.de

Göttingen, March 9, 2021

Resubmission of our manuscript to *PLOS Computational Biology*

Dear editors, dear reviewers,

we thank you for your editorial consideration and the very helpful comments. Please find enclosed a point-to-point response to the reviewer's comments, and a manuscript file with changes highlighted in colour.

In brief, the main improvements in the revised version of the manuscript comprise (1) a clarification of our approach with respect to previous approaches that quantify temporal dependence in neural spike trains, (2) we compare our approach to classical measures, and show more simulated example neurons to illustrate the properties of the new approach, and (3) we revised the definition of the timescale.

In more detail, first, we carve out more clearly that our measure of history dependence assesses the window over which unique predictive information is accumulated; in contrast to e.g. autocorrelation, which assesses how long—potentially redundant—past information can still be read out. Moreover, the conventional estimate of timescale, the autocorrelation time, mixes the effects of strength and timescale of history dependence. In contrast, these are disentangled with our method.

Second, as proposed, we compare the novel measure on the example data sets to other well-established statistics, such as the median interspike interval, the coefficient of variation and the autocorrelation time. Moreover, we demonstrate its properties at a range of simulated model neurons, including the Izhikevich neuron. Third, we replaced the temporal depth of history dependence by a measure of a generalized timescale, which is equivalent to the autocorrelation time, but can also be applied to our measure of history dependence. With its similarity to the autocorrelation time, it facilitates the comparison to past work. In addition, this measure of timescale is more robust to the recording length, and thus further improves quantification.

With grounding our work in a more familiar terrain, and by introducing the robust measure of timescale, we could improve the clarity of our manuscript and method.

We thank you very much for your editorial consideration and are looking forward to your reply,

Lucas Rudelt & Viola Priesemann

Reviewer 1 This paper proposes new metrics for measuring history dependence in neural spike trains, and uses a particular coarse-graining in combination with existing entropy/mutual information estimation methods to estimate this metric for a range of neural spike trains. The authors then try to draw conclusions about their estimated metrics for various real neural spike trains.

The methods aspect of this seems relatively sound. I do have a suggestion for the authors, though, in terms of presentation: I'd put the vast majority of the methods in the Methods rather than the Results section. Basically, the discussion of the curse of dimensionality and the Data Processing Inequality in various forms (large number of bins is curse of dimensionality, can lead to overestimation and small number of bins yields lower MI due to Data Processing Inequality) seem to me to be well-worn statistical ground and not worthy of so much of the Results section.

> Thank you for your summary and your helpful comments. Indeed, we agree with you and had similar discussions during the writing process. However, since the article is aimed at a broad readership that might not be familiar with the issue of over- or underestimation, we found it important to illustrate it here. To incorporate your feedback in the revised manuscript, we included a statement at the beginning of the benchmark results section that encourage readers familiar with the concepts to skip this part (lines 341–344 in the new manuscript).

I'd also emphasize more that your main contribution to estimation of these information quantities is a particularly clever coarse-graining that assumes the recency hypothesis.

> Thank you for this suggestion. We clarified this contribution for the estimation by mentioning it explicitly in the abstract. The relevant passage reads
> "To still account for the vastly different spiking statistics and potentially long history dependence of living neurons, we developed an embedding-optimization approach that does not only vary the number and size, but also an exponential stretching of past bins."
>
> However, we would also like to point out that while our approach is based on established estimators, the way the approach uses them for regularization during the embedding optimization is novel and key to the estimation. As you point out, the coarse-graining with the recency hypothesis is an additional important step, but the approach could be used to optimize any other embedding model.

But that's not my main worry. I'm mainly worried that the metric isn't necessarily the right one for the job. On the chopping block is not just your $R(T)$ (which I would not call a redundancy, but rather just a version of the predictive information divided by H) and T_D (which I have a few comments on later), but also the autocorrelation function (which you discard, for reasons that make sense) and the predictive information (which you essentially have a version of in your numerator, but see Nemenman et al) and all the information measures in ”Anatomy of a Bit” by Ryan James et al.

> Thank you very much for pointing this out. About the predictive information, we regularly refer to it using both terms, predictable or redundant information. It depends on the decoder perspective, whether the information is used or not. In addition, we refer now explicitly to these measures and the additional literature you quote. The relevant passage in the methods summary now reads
> “We quantify history dependence based on the mutual information
>
> $$I(\text{spiking};\text{past}(T)) = H(\text{spiking}) - H(\text{spiking}|\text{past}(T)) \tag{1}$$
>
> between current spiking in a time bin $[t, t + \Delta t)$ and its own past in a past range $[t - T, t)$ (Fig 1B). Here, we assume stationarity and ergodicity, such that the measure is an average over all times t. This mutual information is also called active information storage [5], and is related to the predictive information [18,19]. It quantifies how much of the current spiking information $H(\text{spiking})$ can be predicted from past spiking.”
>
> However, we want to stress that there are two important differences between $R(T)$ and the predictive information: First, $R(T)$ quantifies how well spiking *in the next time bin* can be predicted, similar to the active information storage [5], whereas predictive information also increases the range of predicted spiking with T. Therefore, $R(T)$ can have very distinct behavior as one increases T (for example, the asymptotic rate is zero, see next comment). We chose active information storage over predictive information, because we want to quantify how redundant or predictable the current spiking is, based on its immediate past. From a practical point of view, this quantify is also easier to estimate, because only the past range T has to be embedded.
> Second, we normalize the mutual information by the spiking entropy. This is a crucial step to obtain a measure of statistical dependence, instead of information, similar to the correlation coefficient that normalizes covariance by the variance of the process. See below for more details where we discuss this in light of our novel results.
> Finally, we would like to stress that the main goal here was not to introduce a new information theoretic measure, but to use existing tools from information theory to address a problem that was previously only tackled using measures like the autocorrelation. However, in order to do so, we find it necessary to normalize by the entropy.

Based on my experience playing with these metrics, I’d say the following: – it is likely that T_D will grow with the size of your data set, and so what’s really relevant is the rate of growth; that may be a better way to distinguish between different time series;

> Your are right, the previous measure of temporal depth T_D was highly sensitive to the size of the data set, which we showed in the old S2 and S3 Figs. Therefore, we revised this measure completely. We now define the information timescale τ_R, which is more robust with respect to data size (see new S2 and S3 Figs), and its definition has a nice analogy to the autocorrelation time. However, we feel that you are referring to an asymptotic rate of growth $R(T)/T$ as one lets $T \rightarrow \infty$, similar to the predictive information in [19]. In the case of $R(T)$, this rate will always be zero, because $R(T)$ (with or without normalization) is bounded by one (or the spiking entropy; see previous comment). Thus, no such rate of growth can be defined for this measure.

– it is likely that R(T) has some weird behavior with the time bin size for the present neural patterning that has not yet been discussed and should be;

> We added a supplementary figure (S16 Fig) that shows the dependence of $R(T)$ on the time bin size for the experimental data. While the information timescale τ_R is quite insensitive to the choice of Δt, the total history dependence decreases for small Δt. We added a passage in the methods summary where we discuss and explain our choice of $\Delta t = 5\,\mathrm{ms}$, which reads
> "Finally, all the above measures can depend on the size of the time bin Δt, which discretizes the current spiking activity in time. Too small a time bin holds the risk that noise in the spike emission reduces the overall predictability or history dependence, whereas an overly large time bin holds the risk of destroying coding relevant time information in the neuron's spike train. Thus, we chose the smallest time bin $\Delta t = 5\,\mathrm{ms}$ that does not yet show a decrease in history dependence (S16 Fig)."

– I still have no idea how or if either $R(T)$ or T_D (data set size) capture anything related to history dependence.

> To clarify this, we would like to point you to the new first section in Results, as well as Figs 1, 3, 4, and S14 Fig that clarify the difference between $R(T)$ or τ_R and the autocorrelation time, time-lagged mutual information and the total mutual information (R_{tot} without normalization). For more details see below.

Before I recommend acceptance, I would ask for simulations of an Izhikevich neuron that can adopt different neuron types. The strawmen, in my opinion, should be first the autocorrelation function and then the predictive information. I believe that information measures of time series can reveal the type of neuron or aspects of how it behaves, but I don't see why I should switch from using the predictive information to using $R(T)$ or its relative T_D. What am I getting from $R(T)$ that I'm not getting from predictive information? What is the intuition behind introducing this new measure? What do the authors even mean by "history dependence"? If I am to normalize something like predictive information by single symbol entropy, as the authors do here, what neural spike train do I now correctly classify as having long history dependence that I before believed had little history dependence? As I am missing this intuition from the paper, I cannot recommend acceptance– yet.

> Thank you very much for this comment. First of all, we added the analysis of the Izhikevich neuron, together with the GLIF and a stochastic branching process, as comparison (Fig 4). The history dependence, and the correlation or lagged mutual information clearly show distinct behavior. In addition, we analyzed a binary autoregressive process, where we could control the firing rate via an uncorrelated, external input, as well as the strength and temporal depth of past dependencies in the process (Fig 3). We find that the total history R_{tot} correctly captures an increase in the strength m of past dependencies, whereas the information timescale τ_R is only sensitive to the temporal depth of the process. In contrast, the two aspects are mixed in the autocorrelation time.
>
> The example also addresses your question why the *normalized* mutual information or redundancy $R(T)$ is the right measure for our purpose. The mutual information is proportional to the spiking entropy, which depends crucially on the time bin, as well as the neuron's firing rate. As a consequence, the total mutual information increases strongly with increasing strength of *uncorrelated* inputs, whereas R_{tot} stays almost unaffected, or rather decreases (Fig 3B). Thus, the mutual information cannot clearly distinguish between an increase in input, or history dependence. In addition, we found that the total mutual information is correlated with the firing rates of the neurons, whereas the normalization allows to compare history dependence in neurons with vastly different firing rates (S13 Fig).

Smaller things:
– I would not say that this measure of history dependence has anything to do with the efficient coding hypothesis, which is more about how stimulus is transformed by a neuron so that the neuron has maximal entropy, or sometimes (depending on who's using the term) is about how mutual information between stimulus and neuron is close to the entropy of the neural activity;

> Thank you for this comment. As you point out, there are different formulations of the efficient coding hypothesis. We refer to the first formulation, where a stimulus is transformed by neurons so that they have maximal entropy – here by reducing temporal redundancy within a single spike train. We refer to this line of efficient coding in the introduction when we write “In classical, noise-less efficient coding, history dependence should be low to minimize redundancy and optimize efficiency of neural information transmission [1-3].”
>
> Temporal redundancy is quantified by R_{tot}, such that one can test for signatures of this kind of efficient coding using this measure of history dependence. All

> of this, however, only makes sense if little noise is present, such that the stimulus information is close to the capacity $H(\text{spiking})$ of the neuron. In contrast, when significant noise is present, low history dependence can also be a signature of strong, uncorrelated noise, and cannot be attributed to the efficiency of the encoding. In such a case, additional analyses that assess the noisiness in the stimulus encoding are required. As first hint, we find in an ongoing follow-up project on a data set where neurons are classified as having a significant or no significant receptive field (which could be associated to noisiness of their encoding), that neurons with no significant receptive field actually have higher R_{tot}, consistently across different visual areas (not published yet).

– I would add some words on when your embedding method is likely to fail, which is precisely when initial conditions really really matter and the recency hypothesis is inaccurate– e.g. network of Izhikevich neurons– and which (notably) some might call long-term history dependence.

> We totally agree and mention possible limitations in the discussion:
> "Finally, our approach uses an embedding model that ranges from uniform embedding to an embedding with exponentially stretching past bins—assuming that past information farther into the past requires less temporal resolution. This embedding model might be inappropriate if for example spiking depends on the exact timing of distant past spikes, with gaps in time where past spikes are irrelevant. In such a case, embedding optimization could be used to optimize more complex embedding models that can also account for this kind of spiking statistics."
>
> However, we would like to emphasize that the degree of coarse-graining *is optimized* in our approach, so if the recency hypothesis is inaccurate, a uniform binning will be chosen. If more detailed knowledge about past dependencies is available, more specific embedding models could be optimized using our approach.

If the authors can convince me that their metric $R(T)$ and its relative T_D (which should really be some aspect of how T_D changes with recording length) contain useful information that stumps the predictive information, then I will happily recommend acceptance.

> We hope that with the new figures and clarifications in the text we have convinced you of the usefulness of the analysis using R_{tot} and τ_R, and the differences to predictive information. In addition, for cases where the predictive information is of interest, the embedding optimization approach presented in this paper could facilitate its estimation, as is the case for $R(T)$.

Reviewer 2 This paper is a potentially important contribution to neuroscientific toolbox. The authors propose an extension of existing information theoretic approaches that allows for an unbiased estimation of a neuron's history dependence on temporal depth and history dependence. The paper presents a thorough approach to controlling bias and overfitting. Further, the method is applied to several open datasets and an intriguing finding is described. Finally, the code to apply the methods described in the paper is made available with thorough documentation.

> Thank you for the great summary, your helpful requests and comments and your support for improving the usability of the tool. To summarize our changes, we now extended to link to existing approaches, which will facilitate to put our results into context. We expanded and improved our work, first by introducing a measure of timescale that is technically much closer related to the autocorrelation time, second by extending the analysis to several example model neurons, and finally by including the additional analyses on the experimental data sets that you proposed. We think that now the advantages, the distinction from previous approaches, and also the limitations are now much clearer. In the following, we address each point you raised.

I am enthusiastic but have one minor concern and a few related requests for additional analyses described below. In addition, I made a pull request on Github that may help improve the usability of this tool; hopefully, the authors will build on it to include a few tests of the code. This is not a requirement for this review, but it would be great to see code coverage increase to $> 50\%$.

> Thank you very much for your contribution to the tool! That is really great! Building on your pull request, we have increased testing coverage to 86 %.

The concern is the following. History dependence R depends on the entropy of current spiking conditional on the past, as well as on the entropy of current spiking. The average firing rate of a neuron changes its entropy; presumably, this is the reason that entropy of current spiking is in the denominator. In theory, the product does not depend on the neuron's average firing rate; however, it would be nice to get a demonstration that R_{tot} or T_D do not vary as a function of the GLIF neuron's average firing rate, median ISI, or CV. More importantly, I'd like to see a scatterplot of these quantities vs R_{tot} and T_D in the datasets from Fig. 5.

> We conducted the proposed analyses on the data sets and included them in S13 and S14 Figs. We have also added a paragraph in the results section that analyzes the relation between R_{tot} or τ_R and the median ISI, CV or autocorrelation time. The paragraph reads
> "To better understand how other well-established statistical measures relate to the total history dependence R_{tot} and the information timescale τ_R, we show R_{tot} and τ_R versus the median interspike inteval (ISI), the coefficient of variation

$C_V = \sigma_{\text{ISI}}/\mu_{\text{ISI}}$ of the ISI distribution, and the autocorrelation time τ_C in S14 Fig. Estimates of the total history dependence R_{tot} tend to decrease with the median ISI, and to increase with the coefficient of variation C_V. This result is expected for a measure of history dependence, because a shorter median ISI indicates that spikes tend to occur together, and a higher C_V indicates a deviation from independent Poisson spiking. In contrast, the information timescale τ_R tends to increase with the autocorrelation time, as expected, with no clear relation to the median ISI or the coefficient of variation C_V. However, the correlation between the measures depends on the recorded system. For example in retina ($n = 111$), R_{tot} is significantly anti-correlated with the median ISI (Pearson correlation coefficient: $r = -0.69$, $p < 10^{-5}$) and strongly correlated with the coefficient of variation C_V ($r = 0.90$, $p < 10^{-5}$), and τ_R is significantly correlated with the autocorrelation time τ_C ($r = 0.75$, $p < 10^{-5}$). In contrast, for mouse primary visual cortex ($n = 142$), we found no significant correlations between any of these measures. Thus, the relation between R_{tot} or τ_R and the established measures is not systematic, and therefore one cannot replace the history dependence by any of them."

Regarding the firing rate, we did not find any statistical influence on R_{tot} and τ_R (which replaces T_D) on the data sets (S13 Fig, bottom). In contrast, if one does not normalize by the entropy, one observes an increase in total mutual information with the firing rate (S13 Fig, top) - as expected. We also demonstrate the importance of the normalization in the new Fig 3B, where R_{tot} does not increase as one increases the strength of uncorrelated input, whereas the total mutual information does increase. The relevant passage in the results section reads
"The input strength h increases the firing rate and thus the spiking entropy $H(\text{spiking})$. This leads to a strong increase in the total mutual information $I_{\text{tot}} \equiv \lim_{T\to\infty} I(\text{spiking};\text{past}(T))$, whereas the total history dependence R_{tot} is normalized by the entropy and does slightly decrease (Fig 3B). This slight decrease is expected from a sensible measure of history dependence, because the input is random and has no temporal dependence. In addition, input activations may fall together with internal activations, which slightly reduces the total history dependence."

Note however, that normalizing by the entropy does not mean that R_{tot} will not increase for higher firing rates. As an example, consider the GLIF model neuron, where higher firing rates will result in more past spikes that trigger the spike adaptation. In this case, the total history dependence increases with the rate. Yet, for the GLIF model, it is hard to tune parameters such that one can vary the firing rate, median ISI or CV in a controlled way; hence we did not include such an analysis in the paper.

If authors find no correlation there, it may be instructive to look for a different connection to traditional statistics as described in the first section of Discussion. Surely, we won't find any perfect replacements for history dependence, but if T_D is loosely related to some function of autocorrelation, it will help ground researchers in more familiar terrain.

> This is a great point, and a relation to previous measures of temporal dependence was clearly missing. We added a new introduction Figure (Fig 1) that clarifies the difference between $R(T)$ and measures of temporal dependence such as the autcorrelation $C(T)$ and the lagged mutual information $L(T)$. Moreover, we added two figures (Figs 3 and 4) that clearly show, for three different models, how $R(T)$ capture aspects of history dependence that are not captured by $C(T)$ or $L(T)$. Finally, we added plots that compare the autocorrelation time τ_C to R_{tot} and τ_R in S14 Fig, and added a scatter plot of R_{tot} versus τ_C in the Results section in Fig 7B (previously Fig 5). The relevant passage from the Results section reads
> "Notably, total history dependence and the information timescale varied independently among recorded systems, and studying them in isolation would miss differences between recorded systems, whereas considering them jointly allows to distinguish the different systems. Moreover, no clear differentiation between cortical culture, retina and primary visual cortex is possible using the autocorrelation time τ_C (Fig 7B), with medians $\tau_C \approx 68\,\text{ms}$ (culture), $\tau_C \approx 60\,\text{ms}$ (retina) and $\tau_C \approx 80\,\text{ms}$ (primary visual cortex), respectively."
>
> We also discuss these results in the first section of Discussion, where the relevant paragraph reads
> "A key difference between history dependence $R(T)$ and the autocorrelation or lagged mutual information is that $R(T)$ quantifies statistical dependencies between current spiking and the *entire past spiking* in a past range T (Fig 1B). This has the following advantages as a measure of statistical dependence, and as a footprint of information processing in single neuron spiking. First, $R(T)$ allows to compute the total history dependence, which, from a coding perspective, represents the redundancy of neural spiking with all past spikes; or how much of the past information is also represented when emitting a spike. Second, because past spikes are considered jointly, $R(T)$ captures synergistic effects and dismisses redundant past information (Fig 4). Finally, we found that this enables $R(T)$ to disentangle the strength and timescale of history dependence for the binary autoregressive process. (Fig 3). In contrast, autocorrelation $C(T)$ or lagged mutual information $L(T)$ quantify the statistical dependence of neural spiking on a single past bin with delay T, without considering any of the other bins (Fig 1A). Thereby, they miss synergistic effects; and they quantify redundant past dependencies that vanish once spiking activity in more recent past is taken into account (Fig 4). As a consequence, the timescales of these measures reflect both, the strength and the temporal depth of history dependence in the binary autoregressive process (Fig 3)."

The following lists a few minor suggestions.

In most citations, the name of the journal is missing. Is this by design?

> Thank you very much for your attention, we fixed this issue. The problem was an incompatibility between the biblatex translator and the PLOS template.

In line 78, what does 'discrete past embedding of spiking activity' mean? Do you refer to a 'reduced representation' of the past, or the discrete nature of spiking data? I am trying to discern whether past embedding with binary data has been described in practical terms before.

> The 'discrete' refers to the 'reduced representation' of the past, because, from an information theoretic view, spikes hold an infinite amount of information due to the continuous nature of their time information. We have changed the term to 'binary past embedding', because this is a more precise description of the reduced representation that we use in this paper (even if multiple spikes occur in the same time bin, we represent them by 0 or 1).

In line 152, you may wish to say something like 'while minimizing the risk of overestimation'.

> Thank you for your suggestion, we adopted this formulation in the current manuscript.

Line 163 mentions errorbars, but none are visible in Fig. 2D. I think a different place in the paper mentions 2xSTD errorbars being too small to be visible, but does that come later?

> Thank you for pointing this out. The statement about errorbars not being visible was made in the corresponding results section. To avoid confusion, we included this statement also in the figure caption.

I am confused regarding the status of GLM in this paper. Line 337 justly points out its systematic underestimation of history dependence, while line 194 claims that the authors used GLM as ground truth for $R(T, d, \kappa)$. Please clarify.

> The difference in the two cases is that for the data sets, the model assumptions of the GLM are not met, whereas for the GLIF neuron, they are accurate. Therefore, in case of the GLIF neuron, we use the GLM as an analytical tool to benchmark the model-free estimation approach. On the data sets, however, the underlying model is not known. There, we use the model-free estimates to show that the GLM systematically underestimates history dependence, because the model assumptions do not fully agree with the data. To avoid confusion, we removed the sentence on how the ground truth for the GLIF was computed in the results section, and only refer to Materials and methods. There, we clarify that the GLM only serves as ground truth to this particular model, and not in general. The

> corresponding passage reads
>
> “We can thus 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 $R(T, d, \kappa)$. Note that this is a specific property if this model and does not hold in general. For example in experiments, we found that the GLM accounted for less history dependence than model-free estimates (Fig 6).”

In Fig 4, why are bootstrapping errorbars not centered around the median (bars’ height)?

> The bootstrapping errorbars or $95\,\%$ confidence intervals (here bootstrapped over different sorted units) are not centered around the median, because they do not assume a normal distribution. This is different from errorbars on estimates of $R(T)$, which result from “blocks of blocks” bootstrapping of the time series and assume a normal distribution.

When referring to results from extracellular recordings, it may be best to call the units identified through spikesorting ”single units” rather than ”neurons” to remind us that spikesorting is somewhat subjective.

> This is a great suggestion. As some units are multi units, and others are single units, we now call them all ”sorted units” or simply ”units” throughout the manuscript.

In Fig 5, would it be possible to include a scatterplot of history dependence estimated from GLM?

> The GLM is very costly to optimize, such that it is infeasible to estimate $R(T)$ as a function of T for all the sorted units, which is required to estimate the timescale τ_R. Therefore, we did not include such a scatterplot in Fig 7 (old Fig 5).

Please attempt to interpret the results of Fig 6 further. Why is it that single unit 3 has such a distinctive shape? What might this mean for the corresponding neuron’s information processing? What follow-up would you suggest for researchers using your tool when they see shapes like these? Would inspecting autocorrelograms help? Include any diagnostic information you find helpful.

> Thank you for digging deeper here. We extended the interpretation in the results section and followed your suggestion to add the autocorrelograms to Fig 8 (old Fig 6). The relevant passage in the results reads

> "In particular, sorted units display different signatures of history dependence $R(T)$ as a function of the past range T. For some units, history dependence builds up on short past ranges T (e.g. Fig 8A), for some it only shows for higher T (e.g. Fig 8B), and for some it already saturates for very short T (e.g. Fig 8C). A similar behavior is captured by the autocorrelation $C(T)$ (Fig 8, second row). The rapid saturation in Fig 8C indicates history dependence due to bursty firing, which can also be seen by strong positive correlation with past spikes for short delays T (Fig 8C, bottom). To exclude the effects of different firing modes or refractoriness on the information timescale, we only considered past ranges $T > T_0 = 10\,\mathrm{ms}$ when estimating τ_R, or delays $T > T_0 = 10\,\mathrm{ms}$ when fitting an exponential decay to $C(T)$ to estimate τ_C. The reason is that differences in the integration of past information are expected to show for larger T. This agrees with the observation that timescales among recorded systems were much more similar if one instead sets $T_0 = 0\,\mathrm{ms}$, whereas they showed clear differences for $T_0 = 10\,\mathrm{ms}$ or $T_0 = 20\,\mathrm{ms}$ (S15 Fig)."

Related: What is the interpretation of a peak followed by decay in $R(T)$ as in Fig S7, row 2, middle two?

> This is a great question. A peak as in Fig S7, row 2 is an artefact of the estimation. It arises because the embedding-optimized estimator first captures relevant past dependencies as T increases. For larger T, however, these dependencies cannot be resolved due to the regularization and thus limited number of past bins. In theory, $R(T)$ is monotonously increasing with T, because more past information can only increase the mutual information. We explicitly use this knowledge when estimating R_{tot} and τ_R (see lines 846–866 in Materials and methods), such that this behavior has no negative impact on our key observables.

There is a typo in line 1298 and in caption to S4.

> Thank you, both has been updated.

The sentence that starts on line 1326 is too long. Also, it may be good to italicize 'blocks of blocks' here.

> Thank you, we adapted both.

Reviewer 3

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

• Summary of the paper and novelties This work investigates how to reliably quantify the dependence of a single neuron's spiking on its own preceding activity, called history dependence. Previous studies used limited representations of past activity (the so-called past embedding) to estimate information theory-based measures. Here it is argued that a careful embedding of past activity is crucial. A novel embedding-optimization method is proposed here that optimizes temporal binning of past spiking to capture most of the magnitude and the temporal depth of history dependence. The new method is validated against simulated data of a LIF neuron model and empirical data from different databases that account for a large variety of spiking statistics.
• Strengths The main strengths of the work are:
- It is demonstrated that previous ad hoc embedding strategies are likely to capture much less history dependence, or lead to estimates that severely overestimate the true history dependence. The new method maximizes the estimated history dependence while avoiding overestimation.
- The new method is flexible enough to account for the variety of spiking statistics encountered in experiments.

> Thank you for the great summary and your positive and helpful comments. Below, you clearly pointed out the problems with estimating the temporal depth. Thereby, you stimulated us to come up with a different measures of a timescale. This new measure, which we call information timescale, is not only more robust with respect to the data size, but also allows a much better comparison to the timescale of autocorrelation. We also agree with you that the limitations of the approach should be discussed explicitly in the discussion, and have added additional paragraphs that address the limitations that you pointed out. Below, we address your points one by one.

• Weaknesses and suggestions A weak point of the work is that for spike trains with long temporal depths (e.g., larger than 3 seconds, as in Fig. 3 C), the temporal depth estimated by the optimization method is much smaller (630 ms). This is a critical point to discuss in terms of possible limitations to estimate the timescale of neural processing at different stages of the brain.

> This is a very important point that we now solved by improving our measure of the timescale of history dependence, the information timescale τ_R. This quantity is more robust with respect to the data size (see S2 and S3 Figs), while still resolving the differences in timescale between the data sets (Fig 7). Nonetheless, it remains challenging to estimate the correct timescale τ_R if the true timescale is so large as in the GLIF model neuron, where adaptation effects last up to 22s into the past (although the underestimation is much less than for the temporal depth). We added a paragraph in the discussion about this limitation. The relevant passage reads

> "Moreover, there might be cases where a model-free estimation of the true timescale might be infeasible because of the complexity of past dependencies (S2 Fig, neuron with a 22 seconds past kernel). In this case, only $\approx 80\,\%$ of the true timescale could be estimated on a 90 minute recording."
>
> However, to demonstrate that the method can in principle estimate the true timescale, we replaced the results on the GLIF model with 22s kernel with results on a truncated version of the adaptation kernel with 1s kernel (Fig 5), and moved the previous results to Supplementary information (S1 and S2 Figs).

Another drawback of the new optimization methods is that they perform worse on short recordings: the estimated history dependence is overestimated when applying BBC to recordings of 3 minutes (S1 Fig) and the estimated temporal depth is underestimated to half of the real temporal depth (S2 Fig). This aspect might be discussed in the paper, analyzing possible limitations on application of optimization techniques to experimental data of short length.

> We totally agree with you. Originally, these limitations were only discussed in the practical guidelines at the end of Methods and Materials section. However, as these limitations are of key relevance to the embedding optimization and analysis of history dependence, we added two paragraphs on these limitations in the discussion. The relevant passages read
>
> "In contrast, the generalized timescale can be directly applied to estimates of the history dependence $R(T)$ to yield the information timescale τ_R without any further assumptions or fitting models. However, we found that estimates of τ_R can depend strongly on the estimation method and embedding dimension (S12 Fig) and the size of the data set (S2 and S3 Figs). The dependence on data size is not so strong for the practical approach of optimizing up to $d_{\text{max}} = 5$ past bins, but still we recommend to use data sets of similar length when aiming for comparability across experiments."
>
> and
>
> "Another downside of quantifying the history dependence $R(T)$ is that its estimation requires more data than fitting the autocorrelation time τ_C. To make best use of the limited data, we here devised the embedding optimization approach that allows to find the most efficient representation of past spiking for the estimation of history dependence. Even so, we found empirically that a minimum of 10 minutes of recorded spiking activity are advisable to achieve a meaningful quantification of history dependence and its timescale (S2 and S3 Figs). In addition, for shorter recordings, the analysis can lead to mild overestimation due to over-optimizing embedding parameters on noisy estimates (S1 Fig). This overestimation can, however, be avoided by cross-validation, which we find to be particularly relevant for the Bayesian bias criterion (BBC) estimator."

> Regarding the underestimation of the temporal depth, we would like to point out that the information timescale that we introduce in the revised version is more robust to underestimation (new S2 Fig).

Some minor suggestions:
- Line 60: Could you comment on why the time bin of current spiking is chosen to be 5 ms?

> This is an important question, and we have added a part to the Methods summary that explains how we chose the time bin, which we also support by a comparison of the experimental results for different choices of time bins (S16 Fig). The relevant passage reads
> "Finally, all the above measures can depend on the size of the time bin Δt, which discretizes the current spiking activity in time. Too small a time bin holds the risk that noise in the spike emission reduces the overall predictability or history dependence, whereas an overly large time bin holds the risk of destroying coding relevant time information in the neuron's spike train. Thus, we chose the smallest time bin $\Delta t = 5\,\mathrm{ms}$ that does not yet show a decrease in history dependence (S16 Fig)."

- Fig. 1 it is included in the Methods summary but is not well described in the text. Either move it to Methods, or further explain it here. In the figure caption, please provide more details of the figure, e.g., explain what is ML, NSB and BBC.

> Thank you. We moved the figure to the Methods (now Figure 10), and expanded the figure caption.

- Fig S1 and paragraph between lines 272 and 286: how is each half of the data selected for cross-validation? Are multiple rounds of cross-validation performed using different partitions (in this case different halves) of the data?

> We chose the most simple solution and literally take the first half of the data for the optimization of embedding parameters, and the second half for the optimization. Only one round of cross-validation is performed. What matters is that the set of embedding parameters is optimized on a different data set than the data set that is used to estimate $R(T)$. We edited the results paragraph and the figure caption to make this point more clear.

- Fig 4C: why BBC is computed with d = 20, and shuffling with d = 5?

> We agree that this selection of estimates might be confusing, and have added Shuffling with $d_{\max} = 20$ to Fig 6C (old Fig 4C). Now, all estimates from Fig 6B,D that allow exponential embedding are shown.

- Fig S4 is not referenced in the text.

> Thank you for pointing this out, S4 and S5 Figs are now referenced in the results section on the benchmark model in lines 350 and 424.

- Typo: “errorbars” instead of “error bars” (for example, in line 262).

    Thank you, this was fixed.

- The publication year is missing in references.

    Thank you for your attention, the issue is fixed in the current version of the manuscript.

Methods are written in an appropriate and informative way

The paper is well written and concepts are provided in a correct, clear and suitable way.

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
E-mail: lucas.rudelt@ds.mpg.de (LR), viola.priesemann@ds.mpg.de (VP)

Abstract

Information processing can leave distinct footprints on the statistics of neural spiking. For example, efficient coding minimizes the statistical dependencies on the spiking history, while temporal integration of information may require the maintenance of information over different timescales. To investigate these footprints, we developed a novel approach to quantify history dependence within the spiking of a single neuron, using the mutual information between the entire past and current spiking. This measure captures how much past information is necessary to predict current spiking. In contrast, classical time-lagged measures of temporal dependence like the autocorrelation capture how long—potentially redundant—past information can still be read out. Strikingly, we find for model neurons that our method disentangles the *strength* and *timescale* of history dependence, whereas the two are mixed in classical approaches. When applying the method to experimental data, which are necessarily of limited size, a reliable estimation of mutual information is only possible for a coarse temporal binning of past spiking, a so called past embedding. To still account for the vastly different spiking statistics and potentially long history dependence of living neurons, we developed an embedding-optimization approach that does not only vary the number and size, but also an exponential stretching of past bins. For extra-cellular spike recordings, we found that the strength and timescale of history dependence indeed can vary independently across experimental preparations. While hippocampus indicated strong and long history dependence, in visual cortex it was weak and short, while in vitro the history dependence was strong but short. This work enables an information theoretic characterization of history dependence in recorded spike trains, which captures a footprint of information processing that is beyond time-lagged measures of temporal dependence. To facilitate the application of the method, we provide practical guidelines and a toolbox.

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 [1–3]. 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 [5–7] and associative learning [8]. In addition to its magnitude, the timescale of history dependence provides an important footprint of processing at different processing stages in the brain [9–11]. This is because higher-level processing requires integrating information on longer timescales than lower-level processing [12]. Therefore, history dependence in neural spiking should reach further into the past for neurons involved in higher level processing [9, 13]. Quantifying history dependence and its timescale could probe these different footprints and thus yield valuable insights on how neural coding and information processing is organized in the brain.

Often, history dependence is characterized by how much spiking is correlated with spiking with a certain time lag [14, 15]. From the decay time of this lagged correlation, one obtains an intrinsic timescale of how long past information can still be read out [9–11, 16]. However, to quantify not only a timescale of statistical dependence, but also its strength, one has to quantify how much of a neuron's spiking depends on its *entire past*. Here, this is done with the mutual information between the spiking of a neuron and its own past [17], also called active information storage [5–7], or predictive information [18, 19].

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 [20, 21], 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 [22–26]. The downside is that this requires a lot of data, otherwise estimates can be severely biased [27, 28]. A lot of work has been devoted to finding less biased estimates, either by correcting bias [28–31], or by using Bayesian inference [32–34]. 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 [35]. For example, many studies infer history dependence and transfer entropy by embedding the past spiking using a single bin [26, 36].

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 methods summary where we introduce the measure of history dependence and the information timescale, as well as the embedding optimization 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 first compare the measure of history dependence with classical time-lagged measures of temporal dependence on different models of neural spiking activity. Second, 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 the information timescale. For the reader interested in applying the method, we provide practical guidelines in Fig 9 and 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 [37].

Methods summary

Definition of history dependence. First, we define history dependence $R(T)$ in the spiking of a single neuron. We quantify history dependence based on the mutual information

$$I(\text{spiking}; \text{past}(T)) = H(\text{spiking}) - H(\text{spiking}|\text{past}(T)) \tag{1}$$

between current spiking in a time bin $[t, t + \Delta t)$ and its own past in a past range $[t - T, t)$ (Fig 1B). Here, we assume stationarity and ergodicity, such that the measure is an average over all times t. This mutual information is also called active information storage [5], and is related to the predictive information [18, 19]. It quantifies how much of the current spiking information $H(\text{spiking})$ can be predicted from past spiking. The spiking information is given by the Shannon entropy [38]

$$H(\text{spiking}) = -p(\text{spike}) \log_2 p(\text{spike}) - (1 - p(\text{spike})) \log_2 (1 - p(\text{spike})), \tag{2}$$

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 $H(\text{spiking}|\text{past}(T))$ (see Materials and methods) quantifies the average spiking information (in the sense of entropy) that remains when dependencies on past spiking are taken into account. Note that past dependencies can only reduce the average spiking information, i.e. $H(\text{spiking}|\text{past}(T)) \leq H(\text{spiking})$. The difference between the two then gives the amount of spiking information that is redundant or entirely predictable from the past.

To transform this measure of information into a measure of statistical dependence, we normalize the mutual information by the entropy $H(\text{spiking})$ and define history dependence $R(T)$ as

$$R(T) \equiv \frac{I(\text{spiking}; \text{past}(T))}{H(\text{spiking})} = 1 - \frac{H(\text{spiking}|\text{past}(T))}{H(\text{spiking})} \in [0, 1]. \tag{3}$$

While the mutual information quantifies the *amount* of predictable information, $R(T)$ gives the *proportion* of spiking information that is predictable or redundant with past spiking. As such, it interpolates between the following intuitive extreme cases: $R(T) = 0$ corresponds to independent and $R(T) = 1$ to entirely predictable spiking. Moreover, while the entropy and thus the mutual information $I(\text{spiking}; \text{past}(T))$ increases with the firing rate (see S13 Fig for an example on real data), the normalized $R(T)$ is comparable across recordings of neurons with very different firing rates. Finally, all the above measures can depend on the size of the time bin Δt, which discretizes the current spiking activity in time. Too small a time bin holds the risk that noise in the spike emission reduces the overall predictability or history dependence, whereas an overly large time bin holds the risk of destroying coding relevant time information in the neuron's spike train. Thus, we chose the smallest time bin $\Delta t = 5\,\text{ms}$ that does not yet show a decrease in history dependence (S16 Fig).

Fig 1. Illustration of history dependence and related measures in a neural spike train. (A) For the analysis, spiking is represented by 0 or 1 in a small time bin Δt (grey box). Autocorrelation $C(T)$ or the lagged mutual information $L(T)$ quantify the statistical dependence of spiking on past spiking in a single past bin with time lag T_i (green box). (B) In contrast, history dependence $R(T_i)$ quantifies the dependence of spiking on the entire spiking history in a past range T_i. The gain in history dependence $\Delta R(T_i) = R(T_i) - R(T_{i-1})$ quantifies the increase in history dependence by increasing the past range from T_{i-1} to T_i, and is defined in analogy to the lagged measures. (C) Autocorrelation $C(T)$ and lagged mutual information $L(T)$ for a typical example neuron (mouse, primary visual cortex). Both measures decay with increasing T, where $L(T)$ decays slightly faster due to the non-linearity of the mutual information. Timescales τ_C and τ_L (vertical dashed lines) can be computed either by fitting an exponential decay (autocorrelation) or by using the generalized timescale (lagged mutual information). (D) In contrast, history dependence $R(T)$ increases monotonically for systematically increasing past range T, until it saturates at the total history dependence R_{tot}. From $R(T)$, the gain $\Delta R(T_i)$ can be computed between increasing past ranges T_{i-1} and T_i (grey dashed lines). The gain $\Delta R(T)$ decays to zero like the time-lagged measures, with information timescale τ_R (dashed line).

Total history dependence and the information timescale. Here, we introduce measures to quantify the strength and the timescale of history dependence independently. First, note that the history dependence $R(T)$ monotonically increases with the past range T (Fig 1D), until it converges to the *total history dependence*

$$R_{\text{tot}} \equiv \lim_{T \to \infty} R(T). \tag{4}$$

The total history dependence R_{tot} quantifies the proportion of predictable spiking information once the entire past is taken into account.

While the history dependence $R(T)$ is monotonously increasing, the *gain* in history dependence $\Delta R(T_i) \equiv R(T_i) - R(T_{i-1})$ between two past ranges $T_i > T_{i-1}$ tends to decrease, and eventually decreases to zero for $T_i, T_{i-1} \to \infty$ (Fig 1D). This is in analogy to time-lagged measures of temporal dependence such as the autocorrelation $C(T)$ or lagged mutual information $L(T)$ (Fig 1A,C). Moreover, because $R(T)$ is monotonically increasing, the gain cannot be negative, i.e. $\Delta R(T) \geq 0$. From $\Delta R(T_i)$, we quantify a characteristic timescale τ_R of history dependence similar to an autocorrelation time. In analogy to the integrated autocorrelation time [39], we define the *generalized timescale*

$$\tau_R \equiv \sum_{i=1}^{n} \bar{T}_i \frac{\Delta R(T_i)}{\sum_{j=1}^{n} \Delta R(T_j)} - T_0. \tag{5}$$

as the average of past ranges $\bar{T}_i = (T_i + T_{i-1})/2$, weighted with their gain $\Delta R(T_i) = R(T_i) - R(T_{i-1})$. Here, steps between two past ranges T_{i-1} and T_i should be chosen small enough, and summing the middle points $\bar{T}_i$ of the steps further reduces the error of discretization. T_0 is the starting point, i.e. is the first past range for which $R(T)$ is computed, and was set to $T_0 = 10\,\text{ms}$ to exclude short-term past dependencies like refractoriness (see Materials and methods for details). Moreover, the last past range T_n has to be high enough such that $R(T_n)$ has converged, i.e. $R(T_n) = R_{\text{tot}}$. Here, we set $T_n = 5\,\text{s}$ unless stated otherwise.

To illustrate the analogy to the autocorrelation time, we note that if the gain decays exponentially, i.e. $\Delta R(T_i) \propto \exp\left(-\frac{T_i}{\tau_{\text{auto}}}\right)$ with decay constant τ_{auto}, then $\tau_R = \tau_{\text{auto}}$ for $n \to \infty$ and sufficiently small steps $T_i - T_{i-1}$. The advantage of τ_R is that it also generalizes to cases where the decay is not exponential. Furthermore, it can be applied to any other measure of temporal dependence (e.g. the lagged mutual information) as long as the sum in Eq (5) remains finite, and the coefficients are non-negative. Note that *estimates* of $\Delta R(T_i)$ can also be negative, so we included corrections to allow a sensible estimation of τ_R (Materials and methods). Finally, since τ_R quantifies the timescale over which unique predictive information is accumulated, we refer to it as the *information timescale.*

Binary 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 [35]. 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.

Fig 2. Illustration of embedding optimization to estimate history dependence and the information timescale. (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 farther 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 $\hat{\tau}_R$ and $\hat{R}_{\mathrm{tot}}$ of the information timescale τ_R and total history dependence R_{tot} (vertical and horizontal dashed lines). To compute $\hat{R}_{\mathrm{tot}}$ we average estimates $\hat{R}(T)$ in an interval $[T_D, T_{\max}]$, for which estimates $\hat{R}(T)$ reach a plateau (vertical blue bars, see Materials and methods). For high past ranges T, estimates $\hat{R}(T)$ may decrease because a reliable estimation requires past embeddings with reduced temporal resolution.

Estimation of history dependence with binary 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 [28,30]. 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 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 [31] 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 while minimizing the risk of overestimation. See Materials and methods for details on how we used embedding optimized estimates $\hat{R}(T)$ to compute estimates $\hat{R}_{\text{tot}}$ and $\hat{\tau}_R$ of the total history dependence and information timescale (Fig 2, blue dashed lines).

Results

In the first part, we demonstrate the differences between history dependence and classical measures of temporal dependence for several models of neural spiking activity. We then benchmark the estimation of history dependence using embedding optimization on a tractable neuron model with long-lasting spike adaptation. Moreover, we compare the embedding optimization approach to existing estimation methods on a variety of extra-cellular spike recordings. In the last part, we apply this to analyze history dependence for a variety of recorded systems, and compare the results to the autocorrelation and other statistical measures on the data.

Differences between history dependence and time-lagged measures of temporal dependence

The history dependence $R(T)$ quantifies how predictable neural spiking is, given activity in a certain past range T. In contrast, time-lagged measures of temporal dependence like the autocorrelation $C(T)$ [40] or lagged mutual information $L(T)$ [41, 42] quantify the dependence of spiking on activity in a single past bin with delay T (Fig 1A,C; Materials and methods). In the following, we showcase the main differences between the two approaches.

History dependence disentangles the effects of input activation, reactivation and temporal depth of a binary autoregressive process. To show the behavior of the measures in a well controlled setup, we analyzed a simple binary autoregressive process with varying temporal depth l (Fig 3). The process evolves in discrete time steps, and has an active (1) or inactive (0) state (Fig 3A). Active states are evoked either by external input with probability h, or by internal reactivations that are triggered by activity within the past l steps. Each past activation increases the reactivation probability by m, which regulates the strength of history dependence in the process. In the following, we describe how the measures behave as we vary each of the different model parameters, and then summarize the key difference between the measures.

Fig 3. History dependence disentangles the effects of input activation, reactivation and temporal depth of a binary autoregressive process. (A) In the binary autoregressive process, the state of the next time step (grey box) is active (one) either because of an input activation with probability h, or because of an internal reactivation. The internal activation is triggered by activity in the past l time steps (green), where each active state increases the activation probability by m. (B) Increasing the input activation probability h increases the total mutual information, although input activations are random and therefore not predictable. Normalizing the total mutual information by the entropy yields the total history dependence, which decreases mildly with h. (C) Autocorrelation $C(T)$, lagged mutual information $L(T)$ and gain in history dependence $\Delta R(T)$ decay differently with the delay T. For $l = 1$ and $m = 0.8$ (top), autocorrelation $C(T)$ decays exponentially with autocorrelation time τ_C, whereas $L(T)$ decays faster due to the non-linearity of the mutual information. $\Delta R(T)$ is non-zero only for delays shorter or equal to the temporal depth of the process, with much shorter timescale τ_R. For $l = 5$, $C(T)$ and $L(T)$ plateau over the temporal depth, and then decay much slower than for $l = 1$. Again, $\Delta R(T)$ is non-zero only within the temporal depth of the process. Parameters m and h were adapted to match the firing rate and total history dependence between $l = 1$ and $l = 5$. (D) When increasing the reactivation probability m for $l = 1$, timescales of time-lagged measures τ_C and τ_L increase. For history dependence, the information timescale τ_R remains constant, but the total history R_{tot} increases. (E) When varying the temporal depth l, all timescales increased. Parameters h and m were adapted to hold the firing rate and R_{tot} constant.

The input strength h increases the firing rate and thus the spiking entropy $H(\text{spiking})$. This leads to a strong increase in the total mutual information $I_{\text{tot}} \equiv \lim_{T\to\infty} I(\text{spiking}; \text{past}(T))$, whereas the total history dependence R_{tot} is normalized by the entropy and does slightly decrease (Fig 3B). This slight decrease is expected from a sensible measure of history dependence, because the input is random and has no temporal dependence. In addition, input activations may fall together with internal activations, which slightly reduces the total history dependence.

In contrast, the total history dependence R_{tot} increases with the reactivation probability m, as expected (Fig 3D). For the autocorrelation, the reactivation probability m not only influences the magnitude of the correlation coefficients, but also the decay of the coefficients. For autoregressive processes (and $l = 1$), autocorrelation coefficients $C(T)$ decay exponentially [14] (Fig 3C), where the autocorrelation time $\tau_C = -\Delta t/\log(m)$ increases with m and diverges as $m \to 1$ (Fig 3D). The lagged mutual information $L(T)$ is a non-linear measure of time-lagged dependence, and has a very similar behavior as the autocorrelation, with a slightly faster decay and thus smaller generalized timescale τ_L (Fig 3C,D). Note that we normalized $L(T)$ by the spiking entropy H to make it directly comparable to $\Delta R(T)$. In contrast to the time-lagged measures, the gain in history dependence $\Delta R(T)$ is only non-zero for T smaller or equal to the true temporal depth l of the process (Fig 3C). As a consequence, the information timescale τ_R does not increase with m for fixed l (Fig 3D).

Finally, the temporal depth l controls how far into the past activations depend on their preceding activity. Indeed, we find that the information timescale τ_R increases with l as expected (Fig 3C,E). Similarly, the timescales of the time-lagged measures τ_C and τ_L increase with the temporal depth l. Note that parameters m and h were adapted for each l to keep the firing rate and total history dependence R_{tot} constant, such that differences in the timescale can be unambiguously attributed to the increase in l.

To conclude, history dependence disentangles the effects of input activation, reactivation and temporal depth, which provides a comprehensive characterization of past dependencies in the autoregressive model. This is different from the total mutual information, which lacks the entropy normalization and is sensitive to the firing rate. This is also different from time-lagged measures, whose timescales are sensitive to both, the reactivation probability m *and* the temporal depth l. The confusion of effects in the timescales is rooted in the time-lagged nature of the measures—by quantifying past dependencies out of context, $C(T)$ and $L(T)$ also capture *indirect, redundant* dependencies onto past events. Indirect, redundant dependencies arise from unique dependencies, because past states that are uniquely predictive of future activities were in turn uniquely dependent on their own past. The stronger the unique dependence, the longer the indirect dependencies reach into the past, which increases the timescale of time-lagged measures. In contrast, indirect dependencies do not contribute to the history dependence, because they add no predictive information once more-recent past is taken into account.

History dependence dismisses redundant past dependencies and captures synergistic effects. A key property of history dependence is that it evaluates past dependencies in the light of more recent past. This allows the measure to dismiss indirect, redundant past dependencies and to capture synergistic effects. In three common models of neural spiking activity, we demonstrate how this leads to a substantially different characterization of past dependencies compared to time-lagged measures of temporal dependence.

First, we simulated a subsampled branching process [14], which is a minimal model for activity propagation in neural networks and captures key properties of spiking dynamics in cortex [15]. Similar to the binary autoregressive process, active neurons activate neurons in the next time step with probability m, the so called branching parameter, and are activated externally with some probability h. The process was simulated in time steps of $\Delta t = 4\,\text{ms}$ with a population activity of 500 Hz, which was subsampled to obtain a single spike train with a firing rate of 5 Hz (Fig 4A). Similar to the binary autoregressive process, the autocorrelation decays exponentially with autocorrelation time $\tau_C = -\Delta t/\log(m) = 198\,\text{ms}$, and the lagged mutual information decays slightly faster (Fig 4B). In comparison, the gain in history dependence ΔR decays much faster. When increasing the branching parameter m (for fixed firing rate), the total history dependence increased, as in the autoregressive process (S11 Fig). Strikingly, the timescale τ_R remained constant or even decreased for larger $m > 0.967$ and thus higher autocorrelation time $\tau_C > 120\text{ms}$ (S11 Fig), which is different from the binary autoregressive process. The reason is that the branching process evolves at the population level, whereas history dependence is quantified at the single neuron level. Thereby, history dependence also captures indirect dependencies, because the own spiking history reflects the population activity. The higher the branching parameter m, the more informative past spikes are about the population activity, and the shorter is the timescale τ_R over which all the relevant information about the population activity

Fig 4. History dependence dismisses redundant past dependencies and captures synergistic effects (A,B) Analysis of a subsampled branching process. (A) The population activity was simulated as a branching process ($m = 0.98$) and subsampled to yield the spike train of a single neuron (Materials and methods). (B) Autocorrelation $C(T)$ and lagged mutual information $L(T)$ include redundant dependencies and decay much slower than the gain $\Delta R(T)$, with much longer timescales (vertical dashed lines). (C,D) Analysis of an Izhikevich neuron in chattering mode with constant input and small voltage fluctuations. The neuron fires in regular bursts of activity. (D) Time-lagged measures $C(T)$ and $L(T)$ measure both, intra- ($T < 10\,\mathrm{ms}$) and inter-burst ($T > 10\,\mathrm{ms}$) dependencies, which decay very slowly due to regularity of the firing. The gain $\Delta R(T)$ reflects that most spiking can already be predicted from intra-burst dependencies, whereas inter-burst dependencies are highly redundant. In this case, only $\Delta R(T)$ yields a sensible time scale (blue dashed line). (E,F) Analysis of a generalized leaky integrate and fire neuron with long-lasting adaptation filter ξ [3, 43] and constant input. Figure adapted from [44]. (F) Here, $\Delta R(T)$ decays slower to zero than the autocorrelation $C(T)$, and is higher than $L(T)$ for long delays T. Therefore, the dependence on past spikes is stronger when taking more recent past spikes into account ($\Delta R(T)$), as when considering them independently ($L(T)$). Due to these synergistic past dependencies, $\Delta R(T)$ is the only measure that captures the long-range nature of the spike adaptation.

can be collected. Thus, for the branching process, the total history dependence R_{tot} captures the influence of the branching parameter, whereas the information timescale τ_R behaves very differently from the timescales of time-lagged measures.

Second, we demonstrate the difference of history dependence to time-lagged measures on an Izhikevich neuron, which is a flexible model that can produce different neural firing patterns similar to those observed for real neurons [45]. Here, parameters were chosen according to the "chattering mode" [45], with constant input and small voltage fluctuations (Materials and methods). The neuron fires in regular bursts of activity, with consistent timing between spikes within and between bursts (Fig 4C). While time-lagged measures capture all the regularities in spiking and oscillate with the bursts of activity, history dependence correctly captures that spiking can almost be entirely predicted from intra-burst dependencies alone (Fig 4D). History dependence dismisses the redundant inter-burst dependencies and thereby yields a sensible measure of a timescale (blue dashed line).

Finally, we analyzed a generalized leaky integrate-and-fire neuron with long-range spike adaptation (22 seconds) (Fig 4E), which reproduces spike-frequency adaptation as observed for real layer 2/3 pyramidal neurons [3, 43]. For this model, time-lagged measures $C(T)$ and $L(T)$ actually decay to zero much faster than the gain in history dependence $\Delta R(T)$, which is the only measure that captures the long-range adaptation effects of the model (Fig 4F). This shows that past dependencies in this model include synergistic effects, where the dependence is stronger in the context of more recent spikes. This is most likely due to the non-linearity of the model, where past spikes cause a different adaptation when taken together as when considered as the sum of their contributions.

Thus, due to its ability to dismiss redundant past dependencies and to capture synergistic effects, history dependence really provides a complementary characterization of past dependencies compared to time-lagged measures. Importantly, because the approach better disentangles the effects of timescale and total history dependence, the results remain interpretable for very different models, and provide a more comprehensive view on past dependencies.

Embedding optimization captures history dependence for a neuron model with long-lasting spike adaptation

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. Readers that are familiar with the bias problem of mutual information estimation might want to jump to the next part, where we validate that embedding-optimized estimates indeed capture the model's true history dependence, while being robust to systematic overestimation. As a model we chose a generalized leaky integrate-and-fire (GLIF) model with spike frequency adaptation, whose parameters were fitted to experimental data [3, 43]. The model was chosen, because it is equipped with a long-lasting spike adaptation mechanism, and its total history dependence R_{tot} can be directly computed from sufficiently long simulations (Materials and methods). For demonstration, we show results on a variant of the model where adaptation reaches one second into the past, and show results on the original model with a 22 second kernel in S1, S2 and S5 Figs. For simulation, the neuron was driven with a constant input current to achieve an average firing rate of 4 Hz. In the following, estimates $\hat{R}(T)$ are shown for a simulated recording of 90 minutes, whereas the true values $R(T)$ were computed on a 900 minute recording (Materials and methods).

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 $\tau = 20\,\mathrm{ms}$ (i.e. $\kappa = 0$ and $T = d \cdot \tau$). We compared estimates $\hat{R}(\tau, d)$ obtained by maximum likelihood (ML) estimation [28], or Bayesian estimation using the NSB estimator [33], with the model's true $R(\tau, d)$ (Fig 5A). Both estimators accurately estimate $R(\tau, d)$ for up to $d \approx 20$ 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 5A, 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 [31]. The Shuffling estimator is based on the ML estimator and applies a bias correction term (Fig 5B). 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 5B, 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

Fig 5. Embedding optimization captures history dependence for a neuron model with long-lasting spike adaptation. Results are shown for a generalized leaky integrate-and-fire (GLIF) model with long-lasting spike frequency adaptation [3, 43] with a temporal depth of one second (Methods and material). (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 $\tau = 20\,\mathrm{ms}$. Maximum likelihood (ML) [28] and Bayesian (NSB) [33] 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 obtained from 50 repeated simulations, which are very small and thus hardly visible. (C) Embedding optimized BBC estimates $\hat{R}(T)$ (red line) yield accurate estimates of the model neuron's true history dependence $R(T)$, total history dependence R_{tot} and information timescale τ_R (horizontal and vertical dashed lines). The zoom-in (right panel) shows robustness of both regularization methods: For all T the model neuron's $R(T, d^*, \kappa^*)$ lies within errorbars (BBC), or consistently above the Shuffling estimator that provides a lower bound. Here, the model's $R(T, d^*, \kappa^*)$ was computed for the optimized embedding parameters d^*, κ^* that were selected via BBC or Shuffling, respectively (dashed lines). Shaded areas indicate ± two standard deviations obtained by bootstrapping, and colored vertical bars indicate past ranges over which estimates $\hat{R}(T)$ were averaged to compute $\hat{R}_{\mathrm{tot}}$ (Materials and methods).

parameters can be safely optimized to yield better estimates of history dependence.

Optimized embeddings capture the model's true history dependence. 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 $R(T)$. We find that embedding-optimized BBC estimates $\hat{R}(T)$ agree well with the true $R(T)$, such that the model's total history dependence R_{tot} and information timescale τ_R are well estimated (Fig 5C, vertical and horizontal dashed lines). In contrast, the Shuffling estimator underestimates the true $R(T)$ for past ranges $T > 200\,\mathrm{ms}$, such that the model's R_{tot} and τ_R are underestimated (blue dashed lines). For large past ranges $T > 1000\,\mathrm{ms}$, estimates $\hat{R}(T)$ of both estimators decrease again, because no additional history dependence is uncovered, whereas the constraint of an unbiased estimation decreases the temporal resolution of the embedding.

Embedding-optimized estimates are robust to overestimation 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 (Fig 5A,B).

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 (grey dotted line, Fig 5C, 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 (S1 and S4 Figs). 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, e.g. by optimizing embedding parameters on the first half, and computing estimates on the second half of the data (S1 Fig). 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 (S1 Fig). 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 ($\approx$ 90 minutes), such that no cross-validation was applied on the experimental data.

Estimates of the information timescale are sensitive to the recording length. Finally, we also tested the impact of the recording length on estimates $\hat{R}_{\text{tot}}$ of the total history dependence as well as estimates $\hat{\tau}_R$ of the information timescale. While on recordings of 3 minutes embedding optimization still estimated $\approx 95\,\%$ of the true R_{tot}, estimates $\hat{\tau}_R$ were only $\approx 75\,\%$ of the true τ_R (S2 Fig). Thus, estimates of the information timescale τ_R are more sensitive to the recording length, because they depend on the small additional contributions to $R(T)$ for high past ranges T, which are hard to estimate for short recordings. Therefore, we advice to analyze recordings of similar length to make results on τ_R comparable across experiments. 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 true history dependence. 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 the information timescale in experimental spike recordings.

Embedding optimization is key to estimate long-lasting history dependence in extra-cellular spike recordings

Here, we apply embedding optimization to long spike recordings (≈ 90 minutes) from rat dorsal hippocampus layer CA1 [46,47], salamander retina [48,49] and in vitro recordings of rat cortical culture [50]. 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.

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 $R(T)$ as a function of the past range T (see Fig 6B for an example single unit from hippocampus layer CA1, and S6, S7 and S8 Figs for all analyzed sorted units). 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 $R_{\text{tot}} \approx 40\%$ with a temporal depth of almost a second, and an information timescale of $\tau_R \approx 100\,\text{ms}$ (Fig 6B). This indicates that embedding-optimized estimates capture a substantial part of history dependence also in experimental spike recordings.

Fig 6. Embedding optimization is key to estimate long-lasting history dependence in extra-cellular spike recordings. (A) Example of recorded spiking activity in rat dorsal hippocampus layer CA1. (B) Estimates of history dependence $R(T)$ for various estimators, as well as estimates of the total history dependence R_{tot} and information timescale τ_R (dashed lines) (example single unit 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 found with BBC (violet dot), estimates much lower total history dependence. Shaded areas indicate ± two standard deviations obtained by bootstrapping, and small vertical bars indicate past ranges over which estimates of $R(T)$ were averaged to estimate R_{tot} (Materials and methods). (C) An exponential past embedding is crucial to capture history dependence for high past ranges T. For $T > 100\,\text{ms}$, uniform embeddings strongly underestimate history dependence. Shown is the median of embedding-optimized estimates of $R(T)$ with uniform embeddings, relative to estimates obtained by optimizing exponential embeddings, for BBC with $d_{\max} = 20$ (red) and Shuffling with $d_{\max} = 20$ (blue) and $d_{\max} = 5$ (green). Shaded areas show $95\,\%$ percentiles. Median and percentiles were computed over analyzed sorted units in CA1 ($n = 28$). (D) Comparison of total history dependence R_{tot} for different estimation and embedding techniques for three different experimental recordings. For each sorted unit (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 sorted units (CA1: $n = 28$; retina: $n = 111$; culture: $n = 48$).

Importantly, other common estimation approaches fail to capture the same amount of history dependence (Fig 6B,D). To compare how well the different estimation approaches could capture the total history dependence, we plotted for each so the different estimates of R_{tot} relative to the corresponding BBC estimate (Fig 6D). 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 that was estimated using BBC (Fig 6B, 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 parameters were optimized, we argue that the GLM underestimates history dependence because of its specific model assumptions, i.e. no interactions between past spikes. Moreover, we found that the GLM performs worse than embedding optimization with only five past bins. Therefore, we conclude that for typical experimental spike trains, interactions between past spikes are important, but do not require very high temporal resolution. In the remainder of this paper we use the reduced approach (Shuffling $d_{\max} = 5$) to compare history dependence among different recorded systems.

Increasing bin sizes exponentially is crucial to estimate long-lasting history dependence. To demonstrate this, we plotted embedding-optimized BBC estimates of $R(T)$ using a uniform embedding, i.e. equal bin sizes, relative to estimates obtained with exponential embedding (Fig 6C), both for BBC with $d_{\max} = 20$ (red) and Shuffling with $d_{\max} = 20$ (blue) or $d_{\max} = 5$ (green). For past ranges $T > 100\,\mathrm{ms}$, 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 $T = 1\,\mathrm{s}$, and only around 60 % for $T = 5\,\mathrm{s}$ (median over analyzed sorted units in CA1). Therefore, we argue that an exponential embedding is crucial for estimating long-lasting history dependence.

Together, total history dependence and its timescale show clear differences between recorded systems and individual sorted units

Finally, we present results from diverse extracellular spike recordings that show interesting differences in history dependence between sorted units of different recorded systems. In addition to recordings from rat dorsal hippocampus layer CA1, salamander retina and rat cortical culture, we analyzed sorted units in a recording of mouse primary visual cortex using the novel Neuropixels probe [51]. 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 information timescale (Fig 7A). Sorted units in cortical culture and hippocampus layer CA1 have high total history dependence R_{tot} with median over sorted units of $\approx 24\,\%$ and $\approx 25\,\%$, whereas sorted units in retina and primary visual cortex have typically low R_{tot} of $\approx 11\,\%$ and $\approx 8\,\%$. In terms of the information timescale τ_R, sorted units in hippocampus layer CA1 display much higher τ_R with a median of $\approx 96\,\mathrm{ms}$ than units in cortical culture with median τ_R of $\approx 12\,\mathrm{ms}$. Similarly, sorted units in primary visual cortex have higher τ_R with median of $\approx 37\,\mathrm{ms}$ than units in retina with median of $\approx 23\,\mathrm{ms}$. 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, total history dependence and the information timescale varied independently among recorded systems, and studying them in isolation would miss differences between recorded systems, whereas considering them jointly allows to distinguish the different systems. Moreover, no clear differentiation between cortical culture, retina and primary visual cortex is possible using the autocorrelation time τ_C (Fig 7B), with medians $\tau_C \approx 68\,\mathrm{ms}$ (culture), $\tau_C \approx 60\,\mathrm{ms}$ (retina) and $\tau_C \approx 80\,\mathrm{ms}$ (primary visual cortex), respectively.

Fig 7. Together, total history dependence and its timescale show clear differences between recorded systems. (A) Embedding-optimized Shuffling estimates ($d_{\max} = 5$) of the total history dependence R_{tot} are plotted against the information timescale τ_R for individual sorted units (dots) from four different recorded systems (raster plots show spike trains of different sorted units). No clear relationship between the two quantities is visible. The analysis shows systematic differences between the recorded systems: sorted units in rat cortical culture ($n = 48$) and rat dorsal hippocampus layer CA1 ($n = 28$) have higher median total history dependence than units in salamander retina ($n = 111$) and mouse primary visual cortex ($n = 142$). At the same time, sorted units in cortical culture and retina show smaller timescale than units in primary visual cortex, and much smaller timescale than units in hippocampus layer CA1. Overall, recorded systems are clearly distinguishable when jointly considering the total history dependence and information timescale. (B) Total history dependence R_{tot} versus the autocorrelation time τ_C shows no clear relation between the two quantities, similar to the information timescale τ_R. Also, the autocorrelation time gives the same relation in timescale between retina, primary visual cortex and CA1, whereas the cortical culture has a higher timescale (different order of medians on the x-axis). In general, recorded systems are harder to differentiate in terms of the autocorrelation time τ_C as compared to τ_R. Errorbars indicate median over sorted units and 95 % bootstrapping confidence intervals on the median.

To better understand how other well-established statistical measures relate to the total history dependence R_{tot} and the information timescale τ_R, we show R_{tot} and τ_R versus the median interspike inteval (ISI), the coefficient of variation $C_V = \sigma_{\mathrm{ISI}}/\mu_{\mathrm{ISI}}$ of the ISI distribution, and the autocorrelation time τ_C in S14 Fig. Estimates of the total history dependence R_{tot} tend to decrease with the median ISI, and to increase with the coefficient of variation C_V. This result is expected for a measure of history dependence, because a shorter median ISI indicates that spikes tend to occur together, and a higher C_V indicates a deviation from independent Poisson spiking. In contrast, the information timescale τ_R tends to increase with the autocorrelation time, as expected, with no clear relation to the median ISI or the coefficient of variation C_V. However, the correlation between the measures depends on the recorded system. For example in retina ($n = 111$), R_{tot} is significantly anti-correlated with the median ISI (Pearson correlation coefficient: $r = -0.69$, $p < 10^{-5}$) and strongly correlated with the coefficient of variation C_V ($r = 0.90$, $p < 10^{-5}$), and τ_R is significantly correlated with the autocorrelation time τ_C ($r = 0.75$, $p < 10^{-5}$). In contrast, for mouse primary visual cortex ($n = 142$), we found no significant correlations between any of these measures. Thus, the relation between R_{tot} or τ_R and the established measures is not systematic, and therefore one cannot replace the history dependence by any of them.

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 sorted units in primary visual cortex (Fig 8, see S9 Fig for all analyzed sorted units in primary visual cortex). In particular, sorted units display different signatures of history dependence $R(T)$ as a function of the past range T. For some units, history dependence builds up on short past ranges T (e.g. Fig 8A), for some it only shows for higher T (e.g. Fig 8B), and for some it already saturates for very short T (e.g. Fig 8C). A similar behavior is captured by the autocorrelation $C(T)$ (Fig 8, second row). The rapid saturation in Fig 8C indicates history dependence due to bursty firing, which can also be seen by strong positive correlation with past spikes for short delays T (Fig 8C, bottom). To exclude the effects of different firing modes or refractoriness on the information timescale, we only considered past ranges $T > T_0 = 10\,\mathrm{ms}$ when estimating τ_R, or delays $T > T_0 = 10\,\mathrm{ms}$ when fitting an exponential decay to $C(T)$ to estimate τ_C. The reason is that differences in the integration of past information are expected to show for larger T. This agrees with the observation that timescales among recorded systems were much more similar if one instead sets $T_0 = 0\,\mathrm{ms}$, whereas they showed clear differences for $T_0 = 10\,\mathrm{ms}$ or $T_0 = 20\,\mathrm{ms}$ (S15 Fig).

Fig 8. Distinct signatures of history dependence for different sorted units within mouse primary visual cortex. (Top) Embedding-optimized estimates of $R(T)$ reveal distinct signatures of history dependence for three different sorted units (A,B,C) within a single recorded system (mouse primary visual cortex). In particular, sorted units have similar total history dependence R_{tot}, but differ vastly in the information timescale τ_R (horizontal and vertical dashed lines). Note that for unit C, τ_R is smaller than 5 ms and thus doesn't appear in the plot. Shaded areas indicate ± two standard deviations obtained by bootstrapping, and vertical bars indicate the interval over which estimates of $R(T)$ were averaged to estimate R_{tot} (Materials and methods). Estimates were computed with the Shuffling estimator and $d_{\max} = 5$. (Bottom) Autocorrelograms for the same sorted units (A,B, and C, respectively) roughly show an exponential decay, which was fitted (solid grey line) to estimate the autocorrelation time τ_C (grey dashed line). Similar to the information timescale τ_R, only coefficients for delays larger than $T_0 = 10\,\mathrm{ms}$ were considered during fitting.

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 $T \approx 200$ ms in CA1 (Fig 6B), possibly reflecting phase information in the theta cycles [46, 47]. Thus, the analysis does not only serve to investigate differences in history dependence between recorded systems, but also resolves clear differences between sorted units. This could be used to investigate differences in information processing between different cortical layers, different neuron types or neurons with different receptive field properties.

Overall, our results demonstrate that embedding optimization is powerful enough to reveal clear differences in history dependence between sorted units of different recorded systems, but also between units within the same system. Even more importantly, because units 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 experimental data, we developed a method where the embedding of past spiking is optimized for each individual spike train. 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. Based on our results, we arrived at practical guidelines that are summarized in Fig 9. In the following, we contrast history dependence $R(T)$ with time-lagged measures such as the autocorrelation in more detail, clearly discussing the advantages—but also the limitations of the approach. We then discuss how one can relate estimated history dependence to neural coding and information processing based on the example data sets analyzed in this paper.

1) Embedding optimization: The embedding of past-spiking activity should be individually optimized to each spike train, in order to account for very different spiking statistics. This also applies to other information metrics like transfer entropy [52].

2) Regularization: Estimates have to be reliable lower bounds, otherwise one cannot interpret the results (apply Bayesian bias criterion or Shuffling correction).

3) Exponential embedding: Given the limitations on the number of bins, a non-uniform embedding is required to capture long-lasting dependencies. An exponential embedding with max. 5 bins is typically a good compromise between accuracy and computation speed, and enables embedding optimization for large, highly parallel spike recordings.

4) Data requirements: For practical purpose, spike recordings should be sufficiently long (at least 10 minutes). If several recordings are to be analyzed, these should be of similar length to allow for a meaningful comparison of history dependence and its timescale between recordings.

Fig 9. Practical guidelines for the estimation of history dependence in single neuron spiking activity. More details regarding the individual points can be found at the end of Materials and methods.

Advantages and limitations of history dependence in comparison to the autocorrelation and lagged mutual information. A key difference between history dependence $R(T)$ and the autocorrelation or lagged mutual information is that $R(T)$ quantifies statistical dependencies between current spiking and the *entire past spiking* in a past range T (Fig 1B). This has the following advantages as a measure of statistical dependence, and as a footprint of information processing in single neuron spiking. First, $R(T)$ allows to compute the total history dependence, which, from a coding perspective, represents the redundancy of neural spiking with all past spikes; or how much of the past information is also represented when emitting a spike. Second, because past spikes are considered jointly, $R(T)$ captures synergistic effects and dismisses redundant past information (Fig 4). Finally, we found that this enables $R(T)$ to disentangle the strength and timescale of history dependence for the binary autoregressive process. (Fig 3). In contrast, autocorrelation $C(T)$ or lagged mutual information $L(T)$ quantify the statistical dependence of neural spiking on a single past bin with delay T, without considering any of the other bins (Fig 1A). Thereby, they miss synergistic effects; and they quantify redundant past dependencies that vanish once spiking activity in more recent past is taken into account (Fig 4). As a consequence, the timescales of these measures reflect both, the strength and the temporal depth of history dependence in the binary autoregressive process (Fig 3).

Moreover, technically, the autocorrelation time τ_C depends on fitting exponential decay to coefficients $C(T)$. Computing the autocorrelation time with the generalized timescale is difficult, because coefficients $C(T)$ can be negative, and are too noisy for large delays T. While model fitting is in general more data efficient than the model-free estimation presented here, it can also produce biased and unreliable estimates [16]. Furthermore, when the coefficients do not decay exponentially, a more complex model has to be fitted [53], or the analysis simply cannot be applied. In contrast, the generalized timescale can be directly applied to estimates of the history dependence

$R(T)$ to yield the information timescale τ_R without any further assumptions or fitting models. However, we found that estimates of τ_R can depend strongly on the estimation method and embedding dimension (S12 Fig) and the size of the data set (S2 and S3 Figs). The dependence on data size is not so strong for the practical approach of optimizing up to $d_{\max} = 5$ past bins, but still we recommend to use data sets of similar length when aiming for comparability across experiments. Moreover, there might be cases where a model-free estimation of the true timescale might be infeasible because of the complexity of past dependencies (S2 Fig, neuron with a 22 seconds past kernel). In this case, only $\approx 80\,\%$ of the true timescale could be estimated on a 90 minute recording.

Another downside of quantifying the history dependence $R(T)$ is that its estimation requires more data than fitting the autocorrelation time τ_C. To make best use of the limited data, we here devised the embedding optimization approach that allows to find the most efficient representation of past spiking for the estimation of history dependence. Even so, we found empirically that a minimum of 10 minutes of recorded spiking activity are advisable to achieve a meaningful quantification of history dependence and its timescale (S2 and S3 Figs). In addition, for shorter recordings, the analysis can lead to mild overestimation due to over-optimizing embedding parameters on noisy estimates (S2 Fig). This overestimation can, however, be avoided by cross-validation, which we find to be particularly relevant for the Bayesian bias criterion (BBC) estimator. Finally, our approach uses an embedding model that ranges from uniform embedding to an embedding with exponentially stretching past bins—assuming that past information farther into the past requires less temporal resolution. This embedding model might be inappropriate if for example spiking depends on the exact timing of distant past spikes, with gaps in time where past spikes are irrelevant. In such a case, embedding optimization could be used to optimize more complex embedding models that can also account for this kind of spiking statistics.

Differences in total history dependence and information timescale between data sets agree with ideas from neural coding and hierarchical information processing. First, we found that the total history dependence R_{tot} clearly differs among the experimental data sets. Notably, R_{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, 54] and neural adaptation for temporal whitening as observed in experiments [3, 55]. In contrast, R_{tot} was high for neurons in dorsal hippocampus (layer CA1) and cortical culture. In CA1, the original study [47] 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 R_{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 [56], or to recognize statistical structure for associative learning [8]. In cortical culture, neurons are exclusively driven by recurrent input and exhibit strong bursts in the population activity [57]. 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 R_{tot} does indeed provide a footprint of neural coding and information processing.

Second, we observed that the information timescale τ_R increases from retina ($\approx 23\,\text{ms}$) to primary visual cortex ($\approx 37\,\text{ms}$) to CA1 ($\approx 96\,\text{ms}$), in agreement with the idea of a temporal hierarchy in neural information processing [12]. These results qualitatively agree with similar results obtained for the autocorrelation time of spontaneous activity [9], although the information timescales are overall much smaller than the autocorrelation times. Our results suggest that the hierarchy of intrinsic timescales could also show 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 its strength, as well as its timescale. The approach could be used in future experimental studies to quantify history dependence across a diversity of brain areas, e.g. using the novel Neuropixels probe [58], or even across cortical layers within a single area. To this end we provide a toolbox for Python3 [37]. 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 the information timescale. 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
- $\Delta R(T_i) \equiv R(T_i) - R(T_{i-1})$: gain in history dependence
- τ_R: information timescale or generalized timescale of history dependence $R(T)$
- $L(T) \equiv I(X; X_{-T})$: lagged mutual information with time lag T
- τ_L: generalized timescale of lagged mutual information $L(T)$

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^*
- $\hat{R}_{\text{tot}}$: estimated total history dependence, i.e. average $\hat{R}(T)$ for $T \in [T_D, T_{\max}]$, with interval of saturated estimates $[T_D, T_{\max}]$
- $\hat{\tau}_R$: estimated information timescale

Basic definitions

Definition of history dependence. We quantify 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.

$$R(T) \equiv \frac{I(X, \boldsymbol{X}^{-T})}{H(X)} = 1 - \frac{H(X|\boldsymbol{X}^{-T})}{H(X)}. \tag{6}$$

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 [21, 59], i.e.

$$\begin{aligned} I(X, \boldsymbol{X}^{-T}) &= H(X) - H(X|\boldsymbol{X}^{-T}) && (7) \\ &= \sum_{x\in\{0,1\}} p(x) \log_2 \frac{1}{p(x)} - \sum_{\boldsymbol{x}^{-T}\in\{0,1\}^d} p(x, \boldsymbol{x}^{-T}) \log_2 \frac{1}{p(x|\boldsymbol{x}^{-T})} && (8) \\ &= \sum_{x\in\{0,1\}} \sum_{\boldsymbol{x}^{-T}\in\{0,1\}^d} p(x, \boldsymbol{x}^{-T}) \log_2 \frac{p(x|\boldsymbol{x}^{-T})}{p(x)} && (9) \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \log_2 \frac{p(x_{t_n}|\boldsymbol{x}^{-T}_{t_n})}{p(x_{t_n})}. && (10) \end{aligned}$$

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 $\boldsymbol{x}^{-T}_{t_n}$ 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.

Definition of lagged mutual information. The lagged mutual information $L(T)$ [41] for a stationary neural spike trains is defined as the mutual information between present spiking X and past spiking X_{-T} with delay T, i.e.

$$\begin{aligned} L(T) &\equiv I(X; X_{-T}) && (11) \\ &= \sum_{x\in\{0,1\}} \sum_{x_{-T}\in\{0,1\}} p(x, x_{-T}) \log_2 \frac{p(x|x_{-T})}{p(x)} && (12) \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \log_2 \frac{p(x_{t_n}|x_{t_n-T})}{p(x_{t_n})}. && (13) \end{aligned}$$

Here, $x_{t_n} \in \{0,1\}$ indicates the presence of a spike in a time bin $[t_n, t_n + \Delta t)$ and $x_{t_n-T} \in \{0,1\}$ the presence of a spike in a single past bin $[t_n - T, t_n - T + \Delta t)$ at times $t_n = n\Delta t$ that are shifted by Δt. In analogy to $R(T)$, one can apply the generalized timescale to the lagged mutual information to obtain a timescale τ_L with

$$\tau_L \equiv \sum_{i=1}^{n} \bar{T}_i \frac{L(T_i)}{\sum_{i=j}^{n} L(T_j)} - T_0. \qquad (14)$$

Definition of autocorrelation. The autocorrelation $C(T)$ for a stationary neural spike trains is defined as

$$C(T) = \frac{\mathrm{Cov}[x_{t_n}, x_{t_n-T}]}{\mathrm{Var}[x_{t_n}]} = \frac{\langle x_{t_n} x_{t_n-T}\rangle - \langle x_{t_n}\rangle^2}{\langle x_{t_n}^2\rangle - \langle x_{t_n}\rangle^2} \qquad (15)$$

with delay T and x_{t_n} and x_{t_n-T} as above. For an exponentially decaying autocorrelation $C(T) \propto \exp\left(-\frac{T}{\tau_C}\right)$, τ_C is called *autocorrelation time.*

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

$$\Gamma_T(\theta) : \mathcal{F}_T \to S^d \qquad (16)$$

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 $\tau_j = \tau_1 10^{(j-1)\kappa}$. Here, τ_1 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.

$$p(x_t|\boldsymbol{x}_{t,\theta}^{-T}, \boldsymbol{x}_{t,\nu}^{-T}) = p(x_t|\boldsymbol{x}_{t,\theta}^{-T}) \tag{17}$$

for any other past embedding $\boldsymbol{x}_{t,\nu}^{-T}$. If Eq (17) 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 [60], i.e.

$$D_{KL}\left[p(x_t|\boldsymbol{x}_t^{-T})||p(x_t|\boldsymbol{x}_{t,\theta}^{-T})\right] = \sum_{x_t\in\{0,1\}} p(x_t|\boldsymbol{x}_t^{-T}) \log_2 \frac{p(x_t|\boldsymbol{x}_t^{-T})}{p(x_t|\boldsymbol{x}_{t,\theta}^{-T})} \geq 0, \tag{18}$$

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

$$\begin{aligned}
0 &\leq \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{x_{t_n}\in\{0,1\}} p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T}) \log_2 \frac{p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T})}{p(x_{t_n}|\boldsymbol{x}_{t_n,\theta}^{-T})} && (19)\\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{x_{t_n}\in\{0,1\}} p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T}, \boldsymbol{x}_{t_n,\theta}^{-T}) \log_2 \frac{1}{p(x_{t_n}|\boldsymbol{x}_{t_n,\theta}^{-T})} && (20)\\
&\quad - \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{x_{t_n}\in\{0,1\}} p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T}) \log_2 \frac{1}{p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T})} && (21)\\
&= H(X|\boldsymbol{X}_\theta^{-T}) - H(X|\boldsymbol{X}^{-T}), && (22)
\end{aligned}$$

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.

$$R(T,\theta) \equiv 1 - \frac{H(X|\boldsymbol{X}_\theta^{-T})}{H(X)} \leq 1 - \frac{H(X|\boldsymbol{X}^{-T})}{H(X)} = R(T). \tag{23}$$

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

$$\theta^* = \arg\max_{\theta} \hat{R}(T, \theta). \tag{24}$$

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 (S1 Fig). 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 total history dependence and the information timescale. When estimating history dependence $R(T)$ from data, there are some adjustments required to estimate the total history dependence R_{tot} and the information timescale τ_R.

First, estimates $\hat{R}(T)$ are not guaranteed to converge for large past ranges T, but might decrease due to a reduced resolution of embeddings for higher T (Fig 2D). Thus, we estimated an interval $[T_D, T_{\max}]$ for which estimates have converged. Here, the temporal depth T_D and the upper bound $T_{\max}$ are the first and the last past ranges T for which estimates $\hat{R}(T)$ are within one standard deviation of the highest estimate $\hat{R}_{\max}$, i.e. $\hat{R}(T) \geq \hat{R}_{\max} - \sigma_{\hat{R}_{\max}}$ (Fig 2D, vertical blue bars). The standard deviation was estimated by bootstrapping (see Bootstrap confidence intervals). From this interval, an estimate of the total history dependence $\hat{R}_{\text{tot}}$ is obtained by averaging $\hat{R}(T)$ over past ranges $T \in [T_D, T_{\max}]$ (Fig 2D, vertical dashed blue line).

Second, noisy estimates $\hat{R}(T)$ are not guaranteed to be monotonously increasing, such that increments $\Delta\hat{R}(T)$ can be negative. Moreover, noisy estimates can lead to positive $\Delta\hat{R}(T)$ even though the true $R(T)$ has already converged to R_{tot}. This can have a huge effect on the estimated information timescale $\hat{\tau}_R$ if one simply uses these estimates in Eq (5). To avoid this, we use knowledge about the behavior of the true $R(T)$ when estimating $\Delta R(T)$. In particular, we set estimates $\hat{R}(T)$ equal to the largest previous estimate $\hat{R}(T')$ for $T' < T$ if they fall below it, and equal to $\hat{R}_{\text{tot}}$ if they are larger than $\hat{R}_{\text{tot}}$. This enforces that the estimated gain $\Delta\hat{R}(T) \geq 0$ is non-negative, and excludes spurious gain for high T due to noisy estimates.

Finally, the information timescale τ_R can crucially depend on the choice of the minimum past range T_0 in the sum in Eq (5). A $T_0 > 0$ larger than zero allows to ignore short term effects on the history dependence such as the refractory period or different firing modes, which we found beneficial for resolving differences in the timescale among different recorded systems (S15 Fig). In contrast, if the decay is truly exponential, than τ_R is independent of T_0. In this paper, we chose $T_0 = 10\,\text{ms}$ to exclude short term effects, while also not excluding too much past information.

Workflow. The estimation workflow using embedding optimization is summarized in (Fig 10).

Fig 10. Workflow of past-embedding optimization to estimate history dependence and its temporal depth. 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, either using maximum likelihood (ML), or fully Bayesian estimation (NSB). **4)** Apply regularization, i.e. the Bayesian bias criterion (BBC) or Shuffling bias correction, 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 of $R(T)$. **6)** Repeat the estimation for a set of past ranges T to compute estimates of the information timescale τ_R and the total history dependence R_{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

$$\hat{H}(X) = -r\Delta t \log_2 r\Delta t - (1 - r\Delta t)\log_2(1 - r\Delta t), \tag{25}$$

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

$$H(X|\boldsymbol{X}_\theta^{-T}) = H(X, \boldsymbol{X}_\theta^{-T}) - H(\boldsymbol{X}_\theta^{-T}) \tag{26}$$

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 patterns $a_k \in \{0,1\}^{d+1}$. The entropy $H(X, \boldsymbol{X}_\theta^{-T})$ then reads

$$H(X, \boldsymbol{X}_\theta^{-T}) = H(\boldsymbol{\pi}) = -\sum_{k=1}^{K} \pi_k \log_2 \pi_k. \tag{27}$$

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 time step $[t_n, t_n + \Delta t)$ 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

$$p(\boldsymbol{n}|\boldsymbol{\pi}) \propto \prod_{k=1}^{K} \pi_k^{n_k}. \tag{28}$$

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

$$\hat{H}_{\text{ML}}(X, \boldsymbol{X}_\theta^{-T}) = -\sum_{k=1}^{K} \frac{n_k}{N} \log_2 \frac{n_k}{N} \tag{29}$$

or history dependence

$$\hat{R}_{\text{ML}}(T, \theta) = 1 - \frac{\hat{H}_{\text{ML}}(X, \boldsymbol{X}_\theta^{-T}) - \hat{H}_{\text{ML}}(\boldsymbol{X}_\theta^{-T})}{\hat{H}(X)}. \tag{30}$$

The ML estimator has the right asymptotic properties [28, 61], but is known to underestimate the entropy severely when data is limited [28, 62]. 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

$$\text{Bias}\left[\hat{H}_{\text{ML}}(X, \boldsymbol{X}_\theta^{-T})\right] \leq 0. \tag{31}$$

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 [27, 63]. 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)

$$\hat{H}_{\text{MMSE}}(\boldsymbol{n}) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{n}) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) \frac{p(\boldsymbol{n}|\boldsymbol{\pi}) p(\boldsymbol{\pi})}{\int d\boldsymbol{\pi}' p(\boldsymbol{n}|\boldsymbol{\pi}') p(\boldsymbol{\pi}')}. \tag{32}$$

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 $\boldsymbol{n}$

$$p(\boldsymbol{\pi}|\boldsymbol{n}) = \frac{p(\boldsymbol{n}|\boldsymbol{\pi}) p(\boldsymbol{\pi})}{\int d\boldsymbol{\pi}' p(\boldsymbol{n}|\boldsymbol{\pi}') p(\boldsymbol{\pi}')}. \tag{33}$$

The probability for i.i.d. observations $\boldsymbol{n}$ from an underlying distribution $\boldsymbol{\pi}$ is given by the multinomial distribution in Eq (28).

If the prior $p(\boldsymbol{\pi})$ is a conjugate prior to the multinomial likelihood, then the high dimensional integral of Eq (32) can be evaluated analytically [32]. This is true for a class of priors called Dirichlet priors, and in particular for symmetric Dirichlet priors

$$p(\boldsymbol{\pi}|\beta) \propto \prod_{k=1}^{K} \pi_k^{\beta-1}. \tag{34}$$

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 [64]. 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

$$\xi(\beta) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) p(\boldsymbol{\pi}|\beta) = \psi_0(K\beta + 1) - \psi_0(\beta + 1), \tag{35}$$

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

$$p_{NSB}(\boldsymbol{\pi}) \propto \int d\beta \left| \frac{\partial \xi}{\partial \beta} \right| p(\boldsymbol{\pi}|\beta), \tag{36}$$

$$\frac{\partial \xi}{\partial \beta} = K\psi_1(K\beta + 1) - \psi_1(\beta + 1), \tag{37}$$

that weights Dirichlet priors to be flat with respect to the expected entropy $\xi(\beta)$. Since the variance of this expectation vanishes for $K \gg 1$ [64], 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

$$\hat{H}_{NSB}(\boldsymbol{n}) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) \frac{p(\boldsymbol{n}|\boldsymbol{\pi}) p_{NSB}(\boldsymbol{\pi})}{\int d\boldsymbol{\pi}' p(\boldsymbol{n}|\boldsymbol{\pi}') p_{NSB}(\boldsymbol{\pi}')} \tag{38}$$

$$= \frac{\int d\beta \frac{d\xi}{d\beta}(\beta) \hat{H}(\beta) \rho(\beta, \boldsymbol{n})}{\int d\beta' \frac{d\xi}{d\beta}(\beta') \rho(\beta', \boldsymbol{n})}. \tag{39}$$

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

$$\rho(\beta, \boldsymbol{n}) := \frac{\Gamma(K\beta)}{\Gamma(N + K\beta)} \prod_{i=1}^{K} \frac{\Gamma(n_i + \beta)}{\Gamma(\beta)} \tag{40}$$

$$\propto \int d\boldsymbol{\pi}\, p(\boldsymbol{n}|\boldsymbol{\pi})\, p(\boldsymbol{\pi}|\beta) = p(\boldsymbol{n}|\beta), \tag{41}$$

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

$$\hat{H}(\beta) = \sum_{i=1}^{K} \frac{n_i + \beta}{N + K\beta} [\psi_0(N + K\beta + 1) - \psi_0(n_i + \beta + 1)]. \tag{42}$$

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

$$\hat{R}_{\mathrm{NSB}}(T, \theta) = 1 - \frac{\hat{H}_{\mathrm{NSB}}(X, \boldsymbol{X}_\theta^{-T}) - \hat{H}_{\mathrm{NSB}}(\boldsymbol{X}_\theta^{-T})}{\hat{H}(X)}. \tag{43}$$

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 [37]. 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 [34].

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

$$|\hat{R}_{\mathrm{NSB}}(T, \theta) - \hat{R}_{\mathrm{ML}}(T, \theta)| > p \cdot \hat{R}_{\mathrm{NSB}}(T, \theta). \tag{44}$$

Similarly, we define the BBC estimator

$$\hat{R}_{\mathrm{BBC}}(T, \theta) \equiv \begin{cases} \hat{R}_{\mathrm{NSB}}(T, \theta) & \text{if} \quad \hat{R}_{\mathrm{NSB}}(T, \theta) - \hat{R}_{\mathrm{ML}}(T, \theta) \leq p \cdot \hat{R}_{\mathrm{NSB}}(T, \theta), \\ 0 \quad \text{otherwise}. \end{cases} \tag{45}$$

This estimator is designed to be unbiased, and can thus can be used for embedding optimization in Eq (24). 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 5 and S2 Fig).

Shuffling estimator. The Shuffling estimator was originally proposed in [31] 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

$$\hat{R}_{\mathrm{ML}}(T,\theta) = \frac{1}{\hat{H}(X)}\left(\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta}^{-T}) - \hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta}^{-T}|X)\right) \tag{46}$$

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 [28, 62], we know that

$$\begin{aligned}\mathrm{Bias}[\hat{R}_{\mathrm{ML}}(T,\theta)] &= \mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta}^{-T})] - \mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta}^{-T}|X)] && (47)\\ &\leq -\mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta}^{-T}|X)]. && (48)\end{aligned}$$

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 [31] 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

$$p_{\mathrm{sh}}(\boldsymbol{x}_{\theta}^{-T}|x) = \prod_{j=1}^{d} p(x_{\theta,j}^{-T}|x), \tag{49}$$

and the corresponding entropy is

$$H(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X) = \sum_{j=1}^{d} H(X_{\theta,j}^{-T}|X). \tag{50}$$

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

$$\Delta\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X)] \equiv \hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X) - H(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X) \tag{51}$$

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 [27, 63]

$$\mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta}^{-T}|X)] = -\frac{1}{2N\ln 2}\sum_{x\in\{0,1\}}\left(\tilde{K}(x)-1\right) + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{52}$$

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

$$\mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X)] \lesssim \mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)]. \tag{53}$$

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

$$\hat{H}_{\mathrm{ML,sh}}(\boldsymbol{X}_\theta^{-T}|X) \equiv \hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X) - \Delta\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X), \tag{54}$$

and use this to estimate history dependence

$$\hat{R}_{\mathrm{Shuffling}}(T,\theta) \equiv \frac{1}{\hat{H}(X)}\left(\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}) - \hat{H}_{\mathrm{ML,sh}}(\boldsymbol{X}_\theta^{-T}|X)\right). \tag{55}$$

Because of Eq (48) and Eq (53), we know that this estimator is negatively biased in leading order

$$\hat{R}_{\mathrm{Shuffling}}(T,\theta) \lesssim 0 \tag{56}$$

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 [21]. 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

$$\hat{H}(X|\boldsymbol{X}^{-T},\nu) = \frac{1}{N}\sum_{n=1}^{N}\log_2 p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T},\nu)^{-1} \tag{57}$$

which one can plug into Eq (6) to obtain an estimate of the history dependence. The strong law of large numbers [59] 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

$$\nu^* = \arg\max_{\nu}\sum_{n=1}^{N}\log_2 p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T},\nu). \tag{58}$$

We here consider a generalized linear model (GLM) with exponential link function that has successfully been applied to make predictions in neural spiking data [20] and can be used for the estimation of directed, causal information [21]. In a GLM with past dependencies, the spiking probability at time t is described by the instantaneous rate or conditional intensity function

$$\lambda(t|\boldsymbol{x}_t^{-T},\nu) = \lim_{\delta t \to 0} \frac{p(\hat{t} \in [t, t+\delta t]|\boldsymbol{x}_t^{-T},\nu)}{\delta t}. \tag{59}$$

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

$$p(x_t{=}1|\boldsymbol{x}_t^{-T},\nu) = \frac{\lambda(t|\boldsymbol{x}_t^{-T},\nu)\Delta t}{1+\lambda(t|\boldsymbol{x}_t^{-T},\nu)\Delta t}. \tag{60}$$

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

$$\lambda(t|\boldsymbol{x}_t^{-T},\mu,\boldsymbol{h}) = \exp\left(\mu + \sum_{j=1}^{d} h_j x_{t,j}^{-T}\right), \tag{61}$$

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 [65]. In particular, for given past range T, we choose embedding parameters d^*, κ^* that minimize

$$\mathrm{BIC}(d,\kappa) = (d+1)\log_2 N - 2\mathcal{L}^*(d,\kappa), \tag{62}$$

where N is the number of samples and

$$\mathcal{L}^*(d,\kappa) = \sum_{n=1}^{N} \log_2 p(x_{t_n}|\boldsymbol{x}_{t_n,d,\kappa}^{-T},\mu^*,\boldsymbol{h}^*) \tag{63}$$

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

$$\hat{H}_{\mathrm{GLM}}(X|\boldsymbol{X}_{d^*,\kappa^*}^{-T}) = -\frac{1}{N}\mathcal{L}^*(d^*,\kappa^*), \tag{64}$$

which results in the GLM estimator of history dependence

$$\hat{R}_{\mathrm{GLM}}(T) = 1 - \frac{\hat{H}_{\mathrm{GLM}}(X|\boldsymbol{X}_{d^*,\kappa^*}^{-T})}{\hat{H}(X)}. \tag{65}$$

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 [66]. 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 interspike 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 (S10 Fig). Furthermore, we simulated 100 recordings of the same recording length, and for each computed confidence interval for the past range T with the highest estimated history dependence $R(T)$. 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 (S10 Fig). 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 (S1 Fig). To avoid this type of overestimation, we apply one round of 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 $(x_{t_n}, \boldsymbol{x}^{-T}_{t_n,\theta})$ 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 (S1 Fig). 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 [43].

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

$$C\dot{V} = -g_L(V - V_R) + I_{\text{ext}}(t). \tag{66}$$

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 interspike intervals in the simulation were all larger than 10ms. For constant input current I_{ext}, integrating Eq

(66) yields the membrane potential between two spiking events

$$V(t) = V_\infty + (V_0 - V_\infty)e^{-\gamma(t-\hat{t}_0)}, \tag{67}$$

where $\hat{t}_0$ is the time of the most recent spike, $\gamma = g_L/C$ the inverse membrane timescale and $V_\infty = V_R + I_{\mathrm{ext}}/\gamma$ 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

$$p(\hat{t} \in [t, t+\delta t]) = 1 - \exp\left(\int_t^{t+\delta t} \lambda(s) ds\right) \approx \lambda(t)\delta t. \tag{68}$$

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

$$\lambda(t) = \lambda_0 \exp\left(\frac{V(t) - V_T(t)}{\Delta V}\right). \tag{69}$$

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

$$V_T(t) = V_T^* + \sum_{\hat{t}_j < t} \xi(t - \hat{t}_j). \tag{70}$$

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 [43], the following functional form for the adaptation filter was extracted:

$$\xi(s) = \begin{cases} a_\xi & \text{, if } 0 < s \leq T_\xi \\ a_\xi \left(\frac{s}{T_\xi}\right)^{-\beta_\xi} & \text{, if } T_\xi < s < 22\,\mathrm{s}. \end{cases} \tag{71}$$

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 $T_\xi = 8.3\,\mathrm{ms}$, after which it decays like a power-law with exponent $\beta_\xi = 0.93$, until the contributions are set to zero after $22\,\mathrm{s}$.

Model variant with 1s past kernel. For demonstration, we also simulated a variant of the above model with a 1s past kernel

$$\xi^{1\mathrm{s}}(s) = \begin{cases} a_\xi^{1\mathrm{s}} & \text{, if } 0 < s \leq T_\xi \\ a_\xi^{1\mathrm{s}} \left(\frac{s}{T_\xi}\right)^{-\beta_\xi} & \text{, if } T_\xi < s < 1\,\mathrm{s}. \end{cases} \tag{72}$$

All parameters are identical apart from the strength of the kernel $a_\xi^{1\mathrm{s}} = 35.2\,\mathrm{mV}$, which was adapted to maintain a firing rate of $4\,\mathrm{Hz}$ despite the shorter kernel.

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 (66) 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 (68). 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.

Table 1. Simulation parameters of the GLIF-ξ model.

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

The parameters were originally extracted from experimental recordings of (n=14) L5 pyramidal neurons [43].

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 (68) 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

$$p_{\text{sim}}(x_t{=}0|\boldsymbol{x}_t^{-\infty}) = \prod_{j=1}^{\Delta t/\delta t} [1 - \tilde{\lambda}(t + (j-1)\delta t)\delta t], \tag{73}$$

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 $\widetilde{\lambda(t)}$ is computed as $\lambda(t)$ in Eq (69), 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 $p_{\text{sim}}(x_t{=}0|\boldsymbol{x}_t^{-\infty})$.

For sufficiently long simulations, one can make use of the SLLN to compute the conditional entropy

$$H_{\text{sim}}(X|\boldsymbol{X}^{-\infty}) = -\frac{1}{N}\sum_{n=1}^{N} \log_2 p_{\text{sim}}(x_{t_n}|\boldsymbol{x}_{t_n}^{-\infty}), \tag{74}$$

and thus the total history dependence

$$R_{\mathrm{tot}} = 1 - \frac{H_{\mathrm{sim}}(X|\boldsymbol{X}^{-\infty})}{\hat{H}(X)}, \tag{75}$$

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 thus 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 $R(T, d, \kappa)$. Note that this is a specific property if this model and does not hold in general. For example in experiments, we found that the GLM accounted for less history dependence than model-free estimates (Fig 6).

To map the model neuron to a GLM, we plug the membrane and threshold dynamics of Eq (67) and Eq (70) into the equation for the firing rate Eq(69), i.e.

$$\lambda(t) = \exp\left(\log \lambda_0 + V_\infty - V_T^* + \sum_{\hat{t}_j < t} \xi(t - \hat{t}_j) + (V_0 - V_\infty) e^{-\gamma(t - \hat{t}_0)}\right). \tag{76}$$

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

$$(V_0 - V_\infty) e^{-\gamma(t - \hat{t}_0)} \approx \sum_{\hat{t}_j < t} (V_0 - V_\infty) e^{-\gamma(t - \hat{t}_j)}, \tag{77}$$

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

$$\lambda(t) = \exp\left(\mu \sum_{j=1}^{d} h_j x_{t,j}^{-}\right), \tag{78}$$

where

$$\mu = \log \lambda_0 + V_\infty - V_T^* \tag{79}$$

$$h_j = \xi(j\delta t) + (V_0 - V_\infty) e^{-\gamma j \delta t}, \tag{80}$$

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

$$\hat{R}_{\mathrm{GLM}}(T, d, \kappa) = 1 - \frac{\hat{H}_{\mathrm{GLM}}(X|\boldsymbol{X}_{d,\kappa}^{-T})}{\hat{H}(X)}. \tag{81}$$

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 81) 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 [46, 47]. Data from mouse primary visual cortex were recorded according to the UK Animals Scientific Procedures Act (1986).

For all recordings, we only analyzed sorted units with firing rates between 0.5 Hz and 10 Hz to exclude the extremes of either inactive units or units 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 $n = 48$ sorted units that satisfied our requirement on the firing rate. More details on the recording procedure can be found in [67], and details on the data set proper can be found in [50].

Salamander retina. Spikes from larval tiger salamander retinal ganglion cells were recorded *in vitro* by extracting the entire retina on an electrode array [68], while a non-repeated natural movie (leaves moving in the wind) was projected onto the retina. The recording had a total length of about $T_{\mathrm{rec}} \approx 82$ min, and we analyzed in total $n = 111$ sorted units that satisfied our requirement on the firing rate. More details on the recording procedure and the data set can be found in [48, 49]. The spike recording as obtained from the Dryad database [48].

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 $T_{\mathrm{rec}} \approx 90$ min. We analyzed in total $n = 28$ sorted units 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 [46, 47]. 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 sorted units were located using the location of the electrode contacts provided in [51], and the Allen Mouse Common Coordinate Framework [69]. We analyzed in total $n = 142$ sorted units from the mouse "Waksman" that belonged to primary visual cortex (irrespective of their layer) and satisfied our requirement on the firing rate. Second, we only selected units 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 [58] and [51].

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 $\Delta\kappa_{\min} = 0.01$. In Table 2 we summarize the relevant parameters that were used for embedding optimization.

Table 2. Parameters used for embedding optimization.

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)

To facilitate reproduction, we added the settings variable names of the parameters as they are used in the toolbox [37].

Details to Fig 3. For Fig 3B, the process was considered for $l = 1$ and an reactivation probability of $m = 0.8$. For $l = 1$, all probabilities can easily be calculated, with marginal probability to be active $p(x_t = 1) = h/(1 - m + mh)$, and conditional probabilities $p(x_t = 1|x_{t-1} = 1) = h + (1 - h)m$ and $p(x_t = 1|x_{t-1} = 0) = h$. From these probabilities, the total mutual information I_{tot} and total history dependence R_{tot} could be directly computed. We then plotted these quantities as a function of h, where values of h were chosen to vary the firing rate between 0.5 and 10 Hz, with a bin size of $\Delta t = 5$ms. For Fig 3C, the binary autoregressive process was simulated for $n = 10^7$ time steps with $m = 0.8$ ($l = 1$), whereas for $l = 5$, m was adapted to yield approximately the same R_{tot} as for $l = 1$. The input activation probability h was chosen to lead to a fixed probability $p(x = 1) \approx 0.025$, corresponding to 5 Hz firing rate with $\Delta t = 5$ms. Autocorrelation $C(T)$ was computed using the MR.estimator toolbox [53], and $\Delta R(T)$ and $L(T)$ were estimated using plugin estimation. For Fig 3D, the same procedures were applied as in Fig 3C, but now m was varied between 0.5 and 0.95, and h was adapted for each m to hold the firing rate fixed at 5 Hz. For Fig 3E, the same procedures were applied as in Fig 3C, but now l was varied between 1 and 10, and h and m were adapted for each l to hold the firing rate fixed at 5 Hz and R_{tot} fixed at the value for $l = 1$ and $m = 0.8$.

Details to Fig 4A,B. The branching process was simulated using the MR.estimator toolbox, with a time step of $\Delta t = 4\,\mathrm{ms}$, population rate of 500 Hz and subsampling probability of 0.01. Thus, the subsampled spike train had a firing rate of $\approx 5\,\mathrm{Hz}$. The branching parameter was set to $m = 0.98$ with analytic autocorrelation time $\tau_C(m) = 198\,\mathrm{ms}$. For a long simulation, autocorrelation $C(T)$ was computed using the MR.estimator toolbox, $L(T)$ using plugin estimation, and $R(T)$ using embedding optimized Shuffling estimator with $d_{\max} = 20$. The generalized timescales τ_R and τ_L were computed with $T_0 = 10\,\mathrm{ms}$.

Details to Fig 4C,D. The Izhikevich model was simulated with the PyNN toolbox [70], with parameters set to the chattering mode ($a = 0.02$, $b = 0.2$, $c = -50$, $d = 2$), simulation time bin $dt = 0.01\,\mathrm{ms}$, and noisy input with mean 0.011 and standard deviation 0.001. For the analysis, a time step of $\Delta t = 1\,\mathrm{ms}$ was chosen. Apart from that, $C(T)$ and $L(T)$ were computed as for Fig 4B. Here, $R(T)$ was computed with BBC and $d_{\max} = 20$, which revealed higher R_{tot} than Shuffling. To compute τ_R, we set $T_0 = 0$.

Details to Fig 4E,F. The GLIF model was simulated as described in Benchmark neuron model (model with 22s past kernel). The analysis time step was $\Delta t = 5\,\mathrm{ms}$. Apart from that, $C(T)$ and $L(T)$ were computed as for Fig 4B. History dependence $R(T)$ was estimated using a GLM as described in Benchmark neuron model. To compute τ_R, we set $T_0 = 10\,\mathrm{ms}$.

Details to Fig 5A,B. In Fig 5A,B, 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 $\tau = 20\,\mathrm{ms}$ and $d \in [1, 60]$. 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 Fig 5C. In Fig 5C, history dependence $R(T)$ was estimated on a 90 minute recording for 57 different values of T in a range $T \in [10\,\mathrm{ms}, 3\,\mathrm{s}]$. Embedding-optimized estimates were computed with up to $d_{\max} = 25$ past bins, and 95% confidence intervals were computed using the standard deviation over $n = 100$ 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 Fig 6. For Fig 6, history dependence $R(T)$ was estimated for 61 different values of T in a range $T \in [10\,\mathrm{ms}, 5\,\mathrm{s}]$. 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 $R(T_D)$ for the temporal depth T_D that was estimated with BBC. To optimize the estimate, we computed GLM estimates of $R(T_D)$ 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 $\hat{R}(T_D)$ as a best estimate for R_{tot}. For Fig 6A, we plotted only spike trains of channels that were identified as single units. For Fig 6B, 95% confidence intervals were computed using the standard deviation over $n = 100$ bootstrapping samples. For Fig 6C, embedding-optimized estimates with uniform embedding ($\kappa = 0$) were computed with $d_{\max} = 20$ (BBC and Shuffling) or $d_{\max} = 5$ (Shuffling). Medians were computed over the $n = 28$ sorted units in CA1.

Details to Figs 7 and 8. For Figs 7 and 8, history dependence was $R(T)$ was estimated for 61 different values of T in a range $T \in [10\,\mathrm{ms}, 5\,\mathrm{s}]$ using the Shuffling estimator with $d_{\max} = 5$. The autocorrelation coefficients $C(T)$ were computed with the MR.Estimator toolbox [53], and the autocorrelation time τ_C was obtained using the `exponential_offset` fitting function. For each recording, we only analyzed the first 40 minutes to have a comparable recording length. For Fig 7, medians of τ_R, τ_C and R_{tot} were computed over all sorted units that were analyzed, and 95% confidence intervals on the medians were obtained by bootstrapping with $n = 10000$ resamples of the median. For Fig 8, 95% confidence intervals were computed using the standard deviation over $n = 100$ 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 [37], 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 7), or even among neurons within a single area (see examples in Fig 8). 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 [20, 21] does not account for higher-order dependencies. We found that the GLM recovered consistently less total history dependence R_{tot} (Fig 6D). 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 $d_{\max} = 5$ past bins with variable exponential scaling κ could account for most of the total history dependence that was estimated with up to $d_{\max} = 20$ bins (Fig 6D). 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 7 and Fig 8) 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 $d_{\max} = 5$ 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 $d_{\max} = 5$ 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 (S1 Fig). 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%) (S1 Fig). 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 (S1 Fig).

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 total history dependence and information timescale across experiments. The recording length affects estimates of the total history dependence R_{tot}, and especially of the information timescale τ_R. This is because more data allows more complex embeddings, such that more history dependence can be captured. Moreover, complex embeddings are particular relevant for long past ranges T. Therefore, if recordings are shorter, smaller $R(T)$ will be estimated for long past ranges T, leading to smaller estimates of τ_R. We found that for shorter recordings, estimates of R_{tot} were roughly the same as for 90 minutes, but estimates of τ_R were considerably smaller (S2 and S3 Figs).

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 τ_R may not be well resolved. Below 10 minutes, we found that estimates of τ_R could be less than half of the value that was estimated for 90 minutes, and also estimates of R_{tot} showed a notable decrease. In addition, all recordings should have comparable length to prevent that differences in history dependence or timescale are due to different recording lengths.

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.

References

1. Barlow HB. Possible Principles Underlying the Transformations of Sensory Messages. In: Rosenblith WA, editor. Sensory Communication. The MIT Press; 2012. 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. Nature Neuroscience. 2013;16(7):942. doi:10.1038/nn.3431.

4. Atick JJ. Could Information Theory Provide an Ecological Theory of Sensory Processing? Network: Computation in Neural Systems. 1992;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. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. 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. Frontiers in Neuroinformatics. 2014;8. doi:10.3389/fninf.2014.00001.

7. Wibral M, Lizier JT, Priesemann V. Bits from Brains for Biologically Inspired Computing. Frontiers in Robotics and AI. 2015;2. doi:10.3389/frobt.2015.00005.

8. Barlow H. Redundancy Reduction Revisited. Network (Bristol, England). 2001;12(3):241–253.

9. Murray JD, Bernacchia A, Freedman DJ, Romo R, Wallis JD, Cai X, et al. A Hierarchy of Intrinsic Timescales across Primate Cortex. Nature Neuroscience. 2014;17(12):1661–1663. doi:10.1038/nn.3862.

10. Cavanagh SE, Towers JP, Wallis JD, Hunt LT, Kennerley SW. Reconciling Persistent and Dynamic Hypotheses of Working Memory Coding in Prefrontal Cortex. Nature Communications. 2018;9(1):3498. doi:10.1038/s41467-018-05873-3.

11. Wasmuht DF, Spaak E, Buschman TJ, Miller EK, Stokes MG. Intrinsic Neuronal Dynamics Predict Distinct Functional Roles during Working Memory. Nature Communications. 2018;9(1):3499. doi:10.1038/s41467-018-05961-4.

12. Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N. A Hierarchy of Temporal Receptive Windows in Human Cortex. Journal of Neuroscience. 2008;28(10):2539–2550. doi:10.1523/JNEUROSCI.5487-07.2008.

13. 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. Frontiers in Systems Neuroscience. 2018;12. doi:10.3389/fnsys.2018.00055.

14. Wilting J, Priesemann V. Inferring Collective Dynamical States from Widely Unobserved Systems. Nature Communications. 2018;9(1):2325. doi:10.1038/s41467-018-04725-4.

15. Wilting J, Priesemann V. Between Perfectly Critical and Fully Irregular: A Reverberating Model Captures and Predicts Cortical Spike Propagation. Cerebral Cortex. 2019;29(6):2759–2770. doi:10.1093/cercor/bhz049.

16. Zeraati R, Engel TA, Levina A. Estimation of Autocorrelation Timescales with Approximate Bayesian Computations. bioRxiv. 2020; p. 2020.08.11.245944. doi:10.1101/2020.08.11.245944.

17. 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.; 2013. p. 1700–1708. Available from: `http://papers.nips.cc/paper/4873-bayesian-entropy-estimation-for-binary-spike-train-data-using-parame` `pdf`.

18. Bialek W, Tishby N. Predictive Information. arXiv:cond-mat/9902341. 1999;.

19. Bialek W, Nemenman I, Tishby N. Predictability, Complexity, and Learning. Neural Computation. 2001;13(11):2409–2463. doi:10.1162/089976601753195969.

20. 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. Nature. 2008;454(7207):995–999. doi:10.1038/nature07140.

21. Quinn CJ, Coleman TP, Kiyavash N, Hatsopoulos NG. Estimating the Directed Information to Infer Causal Relationships in Ensemble Neural Spike Train Recordings. Journal of Computational Neuroscience. 2011;30(1):17–44. doi:10.1007/s10827-010-0247-2.

22. Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W. Entropy and Information in Neural Spike Trains. Physical Review Letters. 1998;80(1):197–200. doi:10.1103/PhysRevLett.80.197.

23. Panzeri S, Treves A, Schultz S, Rolls ET. On Decoding the Responses of a Population of Neurons from Short Time Windows. Neural Computation. 1999;11(7):1553–1577. doi:10.1162/089976699300016142.

24. Brenner N, Strong SP, Koberle R, Bialek W, van Steveninck RRdR. Synergy in a Neural Code. Neural Computation. 2000;12(7):1531–1552. doi:10.1162/089976600300015259.

25. Panzeri S, Schultz SR. A Unified Approach to the Study of Temporal, Correlational, and Rate Coding. Neural Computation. 2001;13(6):1311–1349. doi:10.1162/08997660152002870.

26. Stetter O, Battaglia D, Soriano J, Geisel T. Model-Free Reconstruction of Excitatory Neuronal Connectivity from Calcium Imaging Signals. PLoS computational biology. 2012;8(8):e1002653. doi:10.1371/journal.pcbi.1002653.

27. Panzeri S, Treves A. Analytical Estimates of Limited Sampling Biases in Different Information Measures. Network: Computation in Neural Systems. 1996;7(1):87–107. doi:10.1080/0954898X.1996.11978656.

28. Paninski L. Estimation of Entropy and Mutual Information. Neural Computation. 2003;15(6):1191–1253. doi:10.1162/089976603321780272.

29. Panzeri S, Schultz SR, Treves A, Rolls ET. Correlations and the Encoding of Information in the Nervous System. Proceedings of the Royal Society of London Series B: Biological Sciences. 1999;266(1423):1001–1012. doi:10.1098/rspb.1999.0736.

30. Panzeri S, Senatore R, Montemurro MA, Petersen RS. Correcting for the Sampling Bias Problem in Spike Train Information Measures. Journal of Neurophysiology. 2007;98(3):1064–1072. doi:10.1152/jn.00559.2007.

31. Montemurro MA, Senatore R, Panzeri S. Tight Data-Robust Bounds to Mutual Information Combining Shuffling and Model Selection Techniques. Neural Computation. 2007;19(11):2913–2957. doi:10.1162/neco.2007.19.11.2913.

32. Wolpert DH, Wolf DR. Estimating Functions of Probability Distributions from a Finite Set of Samples. Physical Review E. 1995;52(6):6841–6854. doi:10.1103/PhysRevE.52.6841.

33. Nemenman I, Bialek W, de Ruyter van Steveninck R. Entropy and Information in Neural Spike Trains: Progress on the Sampling Problem. Physical Review E. 2004;69(5):056111. doi:10.1103/PhysRevE.69.056111.

34. Archer E, Park I, Pillow J. Bayesian Entropy Estimation for Countable Discrete Distributions. Journal of Machine Learning Research. 2013;15.

35. 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; 2005. p. 1–46. Available from: `https://www.worldscientific.com/doi/abs/10.1142/9789812567772_0001`.

36. Palmigiano A, Geisel T, Wolf F, Battaglia D. Flexible Information Routing by Transient Synchrony. Nature Neuroscience. 2017;20(7):1014. doi:10.1038/nn.4569.

37. Rudelt L, Marx DG, Wibral M, Priesemann V. History Dependence Estimator; 2020. Zenodo. Available from: `https://github.com/Priesemann-Group/hdestimator`.

38. Shannon CE. A Mathematical Theory of Communication. The Bell System Technical Journal. 1948;27(3):379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

39. Goodman J, Weare J. Ensemble Samplers with Affine Invariance. Communications in Applied Mathematics and Computational Science. 2010;5(1):65–80. doi:10.2140/camcos.2010.5.65.

40. Brockwell PJ, Davis RA. Time Series: Theory and Methods. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag; 1991. Available from: `https://www.springer.com/de/book/9780387974293`.

41. Chapeau-Blondeau F. Autocorrelation versus Entropy-Based Autoinformation for Measuring Dependence in Random Signal. Physica A: Statistical Mechanics and its Applications. 2007;380:1–18. doi:10.1016/j.physa.2007.02.077.

42. Albers DJ, Hripcsak G. Using Time-Delayed Mutual Information to Discover and Interpret Temporal Correlation Structure in Complex Populations. Chaos. 2012;22(1):013111–013111–25. doi:10.1063/1.3675621.

43. 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. Journal of Neurophysiology. 2012;107(6):1756–1775. doi:10.1152/jn.00408.2011.

44. Shlens J. Notes on Generalized Linear Models of Neurons. arXiv:14041999 [cs, q-bio]. 2014;.

45. Izhikevich EM. Simple Model of Spiking Neurons. IEEE Transactions on Neural Networks. 2003;14(6):1569–1572. doi:10.1109/TNN.2003.820440.

46. Mizuseki K, Sirota A, Pastalkova E, Buzsáki G. Multi-Unit Recordings from the Rat Hippocampus Made during Open Field Foraging.; 2009. Available from: `http://crcns.org/data-sets/hc/hc-2`.

47. Mizuseki K, Sirota A, Pastalkova E, Buzsáki G. Theta Oscillations Provide Temporal Windows for Local Circuit Computation in the Entorhinal-Hippocampal Loop. Neuron. 2009;64(2):267–280. doi:10.1016/j.neuron.2009.08.037.

48. 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; 2017. Available from: `http://datadryad.org/stash/dataset/doi:10.5061/dryad.1f1rc`.

49. Prentice JS, Marre O, Ioffe ML, Loback AR, Tkačik G, Ii MJB. Error-Robust Modes of the Retinal Population Code. PLOS Computational Biology. 2016;12(11):e1005148. doi:10.1371/journal.pcbi.1005148.

50. Marom S. MEA Data. 2018;1. doi:10.17632/4ztc7yxngf.1.

51. Stringer C, Pachitariu M, Carandini M, Harris K. Eight-Probe Neuropixels Recordings during Spontaneous Behaviors; 2019. Available from: `https://janelia.figshare.com/articles/dataset/Eight-probe_Neuropixels_recordings_during_spontaneous_behaviors/7739750`.

52. Wibral M, Vicente R, Lindner M. Transfer Entropy in Neuroscience. In: Wibral M, Vicente R, Lizier JT, editors. Directed Information Measures in Neuroscience. Understanding Complex Systems. Berlin, Heidelberg: Springer; 2014. p. 3–36. Available from: `https://doi.org/10.1007/978-3-642-54474-3_1`.

53. 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. arXiv:200703367 [physics, q-bio]. 2020;.

54. Dong DW, Atick JJ. Temporal Decorrelation: A Theory of Lagged and Nonlagged Responses in the Lateral Geniculate Nucleus. Network: Computation in Neural Systems. 1995;6(2):159–178. doi:10.1088/0954-898X$_{620}$03.

55. Wang XJ, Liu Y, Sanchez-Vives MV, McCormick DA. Adaptation and Temporal Decorrelation by Single Neurons in the Primary Visual Cortex. Journal of Neurophysiology. 2003;89(6):3279–3293. doi:10.1152/jn.00242.2003.

56. Moser EI, Kropff E, Moser MB. Place Cells, Grid Cells, and the Brain's Spatial Representation System. Annual Review of Neuroscience. 2008;31(1):69–89. doi:10.1146/annurev.neuro.31.061307.090723.

57. Masquelier T, Deco G. Network Bursting Dynamics in Excitatory Cortical Neuron Cultures Results from the Combination of Different Adaptive Mechanism. PLOS ONE. 2013;8(10):e75824. doi:10.1371/journal.pone.0075824.

58. Stringer C, Pachitariu M, Steinmetz N, Reddy CB, Carandini M, Harris KD. Spontaneous Behaviors Drive Multidimensional, Brainwide Activity. Science. 2019;364(6437). doi:10.1126/science.aav7893.

59. Meyn SP, Tweedie RL. Markov Chains and Stochastic Stability. Communications and Control Engineering. London: Springer-Verlag; 1993. Available from: `//www.springer.com/de/book/9781447132691`.

60. Kullback S, Leibler RA. On Information and Sufficiency. The Annals of Mathematical Statistics. 1951;22(1):79–86. doi:10.1214/aoms/1177729694.

61. Kiefer J, Wolfowitz J. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. The Annals of Mathematical Statistics. 1956;27(4):887–906. doi:10.1214/aoms/1177728066.

62. Antos A, Kontoyiannis I. Convergence Properties of Functional Estimates for Discrete Distributions. Random Structures & Algorithms. 2001;19(3-4):163–193. doi:10.1002/rsa.10019.

63. Treves A, Panzeri S. The Upward Bias in Measures of Information Derived from Limited Data Samples. Neural Computation. 1995;7(2):399–407. doi:10.1162/neco.1995.7.2.399.

64. Nemenman I, Shafee F, Bialek W. Entropy and Inference, Revisited. arXiv:physics/0108025. 2001;.

65. Schwarz G. Estimating the Dimension of a Model. Annals of Statistics. 1978;6(2):461–464. doi:10.1214/aos/1176344136.

66. Davison AC, Hinkley DV. Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press; 1997. Available from: https://www.cambridge.org/core/books/bootstrap-methods-and-their-application/ED2FD043579F27952363566DC09CBD6A.

67. Shahaf G, Marom S. Learning in Networks of Cortical Neurons. Journal of Neuroscience. 2001;21(22):8782–8788. doi:10.1523/JNEUROSCI.21-22-08782.2001.

68. Marre O, Amodei D, Deshmukh N, Sadeghi K, Soo F, Holy TE, et al. Mapping a Complete Neural Population in the Retina. Journal of Neuroscience. 2012;32(43):14859–14873. doi:10.1523/JNEUROSCI.0723-12.2012.

69. 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. Cell. 2020;181(4):936–953.e20. doi:10.1016/j.cell.2020.04.007.

70. Davison AP, Brüderle D, Eppler J, Kremkow J, Muller E, Pecevski D, et al. PyNN: A Common Interface for Neuronal Network Simulators. Frontiers in Neuroinformatics. 2009;2. doi:10.3389/neuro.11.011.2008.

Supporting information

S1 Fig. Embedding optimization leads to mild overfitting for short recordings, which can be avoided by cross-validation. Shown is the relative bias for two versions of the GLIF model with spike adaption, one with 1s and the other with 22s past kernel. The relative bias refers to the relative difference between embedding-optimized estimates $\hat{R}(T, d^*, \kappa^*)$ and the model's true history dependence $R(T, d^*, \kappa^*)$ for the same optimized embedding parameters d^*, κ^*. The relative bias for $\hat{R}_{\text{tot}}$ was computed by first averaging the relative difference $(\hat{R}(T, d^*, \kappa^*) - R(T, d^*, \kappa^*))/R(T, d^*, \kappa^*)$ for $T \in [T_D, T_{\max}]$, 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 with $d_{\max} = 25$. (Left) For BBC, the relative bias for $\hat{R}_{\text{tot}}$ is zero only if recordings are sufficiently long (> 20 minutes for 1s kernel, and ≈ 90 minutes for 22s kernel). 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 one round of cross-validation is applied, i.e. embedding parameters are optimized on the first, and estimates are computed on the second half of the data, the bias is approximately zero for BBC even for short recordings, 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 $n = 10$ (45 or 90 min) or $n = 30$ (≤ 20 min) different simulations.

S2 Fig. For the simulated neuron model, recording length has little effect on the estimated total history dependence, but large impact on the estimated information timescale. (Left) Mean estimated total history dependence $\hat{R}_{\text{tot}}$ for different recording lengths, relative to the true total history dependence R_{tot} of the model (GLIF with spike adaption with 1s or 22s past kernel). As the recording length decreases, so does $\hat{R}_{\text{tot}}$. However, with only 3 minutes, one does still infer about $\approx 95\%$ of the true R_{tot}. (Right) In contrast, the estimated information timescale $\hat{\tau}_R$ decreases strongly with decreasing recording length. With 3 minutes and less, only $\approx 75\%$ of the true τ_R is estimated on average. Note that for the simpler 1s model (top), an accurate estimation of the true τ_R is possible for 90 minute recordings, whereas for the 22s model (bottom), the estimated $\hat{\tau}_R$ remains below the true value. Shown are mean values for 30 different simulations for T_{rec} between 1 and 20 minutes, and 10 different simulations for 45 and 90 minutes, as well as 95% confidence intervals on the mean based on bootstrapping.

S3 Fig. For experimental data, too, recording length has little effect on estimated total history dependence, but larger impact on the estimated information timescale. (Left) Total history dependence R_{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 estimate about $\approx 95\%$ as much or more of R_{tot} as 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 GLIF model, the estimated information timescale τ_R decreases more strongly with decreasing recording length. With 10 minutes and more, one estimates around $\approx 75\%$ or more of the τ_R that is estimated on a 90 minute recording. Note that for the experimental data, the estimated timescale of the BBC estimator depends more strongly on the recording time, whereas the Shuffling estimator is more robust, especially for $d_{\max} = 5$. Shown is the median with 95% bootstrapping confidence intervals over $n = 10$ randomly chosen sorted units for each recorded system. Before taking the median over sorted units, for each unit 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.

S4 Fig. Example estimation results for the generalized leaky integrate-and-fire model (GLIF) with 1s past kernel. For each recording length, we show the embedding-optimized estimates of history dependence $R(T)$ with and without cross-validation, for BBC (red) and Shuffling (blue) with $d_{\max} = 25$, as well as the ground truth for the same embeddings that were found during optimization (dashed lines). Dashed lines indicate the estimated information timescale $\hat{\tau}_R$ and total history dependence $\hat{R}_{\text{tot}}$. Shaded areas indicate ± two standard deviations obtained by bootstrapping.

S5 Fig. Example estimation results for the generalized leaky integrate-and-fire model (GLIF) with 22s past kernel. For each recording length, we show the embedding-optimized estimates of history dependence $R(T)$ with and without cross-validation, for BBC (red) and Shuffling (blue) with $d_{\max} = 25$, as well as the ground truth for the same embeddings that were found during optimization (dashed lines). Dashed lines indicate the estimated information timescale $\hat{\tau}_R$ and total history dependence $\hat{R}_{\text{tot}}$. Shaded areas indicate ± two standard deviations obtained by bootstrapping.

S6 Fig. Estimation results for all sorted units in rat dorsal hippocampus (layer CA1). For each unit, 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 estimates of the information timescale τ_R and total history dependence R_{tot}. Also shown is the embedding-optimized GLM estimate (violet square) with a past range equal to the temporal depth that was found with the BBC estimator.

S7 Fig. Estimation results for all sorted units in rat cortical culture. For each unit, 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 estimates of the information timescale τ_R and total history dependence R_{tot}. Also shown is the embedding-optimized GLM estimate (violet square) with a past range equal to the temporal depth that was found with the BBC estimator.

S8 Fig. Estimation results for all sorted units in salamander retina. For each unit, 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 estimates of the information timescale τ_R and total history dependence R_{tot}. Also shown is the embedding-optimized GLM estimate (violet square) with a past range equal to the temporal depth that was found with the BBC estimator.

S9 Fig. Estimation results for all sorted units in mouse primary visual cortex. For each unit, we show the embedding-optimized Shuffling estimates of history dependence $R(T)$ for $d_{\max} = 5$. Dashed lines indicate estimates of the information timescale τ_R and total history dependence R_{tot}.

S10 Fig. Bootstrapping yields accurate estimates of standard deviation and confidence intervals. (Left) Shown is the standard deviation on BBC estimates (blue) obtained from 250 "blocks of blocks" bootstrap samples on a single recording (GLIF model with 22s past kernel). It agrees well with the true standard deviation (black), which we estimated from 100 repeated simulations of the same recording length and embedding. 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 $T = T_D$ 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 $\sigma(R)$ have approximately the claimed confidence level (CI accuracy). Standard deviation was estimated from 250 "blocks of blocks" bootstrap samples. For each recording length, we computed estimates $\hat{R}$ and the bootstrapping confidence intervals on the 100 simulations. We then computed the confidence level (CI accuracy) by counting how often the true value of R was contained in the estimated confidence interval (green line). Estimates and the true value of R were computed for the same typical embedding parameters d^*, κ^* and $T = T_D$ as before.

S11 Fig. Total history dependence and information timescale for increasing branching parameter m. Similar to the binary autoregressive process, increasing the branching parameter m increases the total history dependence R_{tot}, whereas the information timescale τ_R stays constant, or even decreases for high m. For each m, the input activation probability h was adapted to hold the firing rate fixed at 5 Hz.

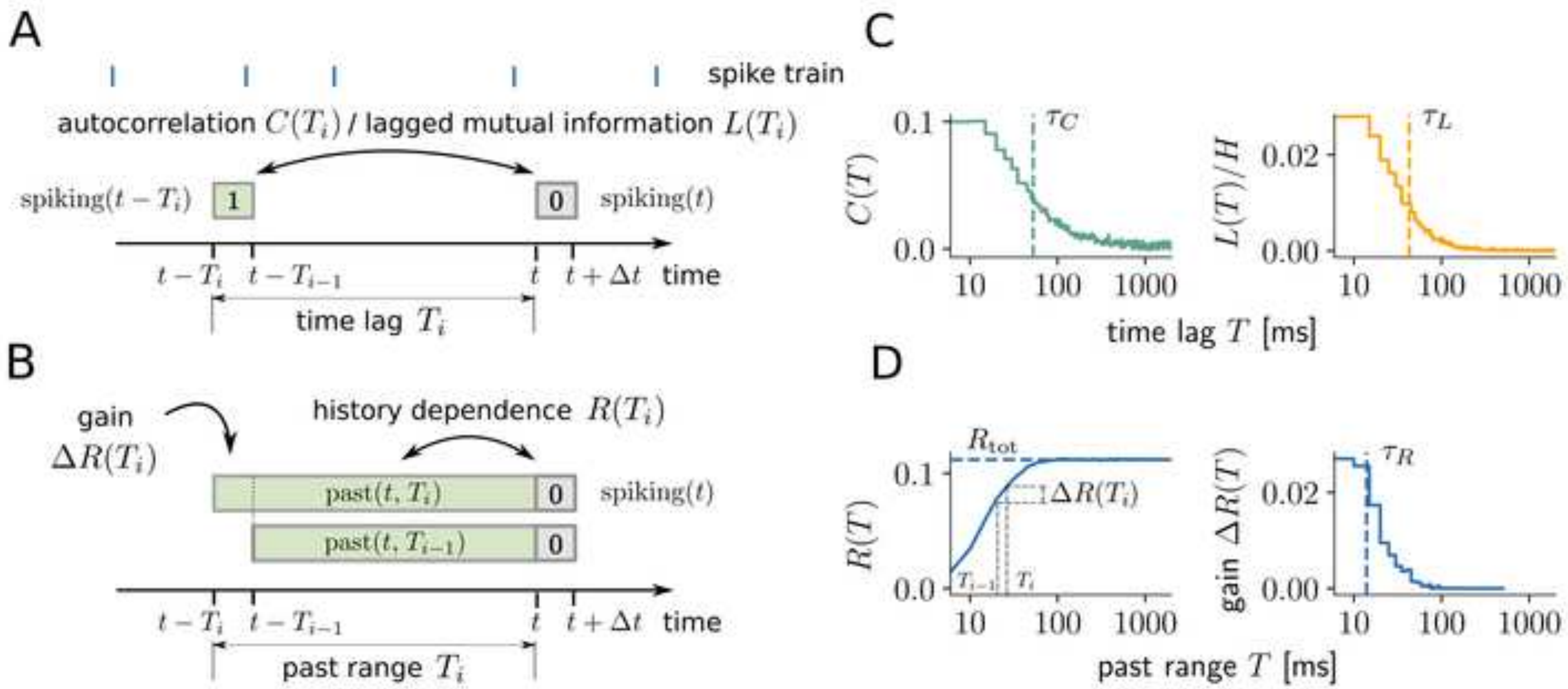
S12 Fig. The estimated information timescale varies between estimators. For each sorted unit (grey dots), estimates of the information timescale τ_R are plotted relative to the corresponding BBC estimate for $d_{\max} = 20$. The BBC estimator tends to estimate higher timescales than the Shuffling estimator on recordings of CA1 and cortical culture, whereas for retina the medians of different estimators are more similar. Although estimates of the timescale are highly variable between estimators, Shuffling with only $d_{\max} = 5$ past bins still estimates timescales of at least 80 % of the timescales that are estimated with BBC. Errorbars indicate median over sorted units and 95 % bootstrapping confidence intervals on the median.

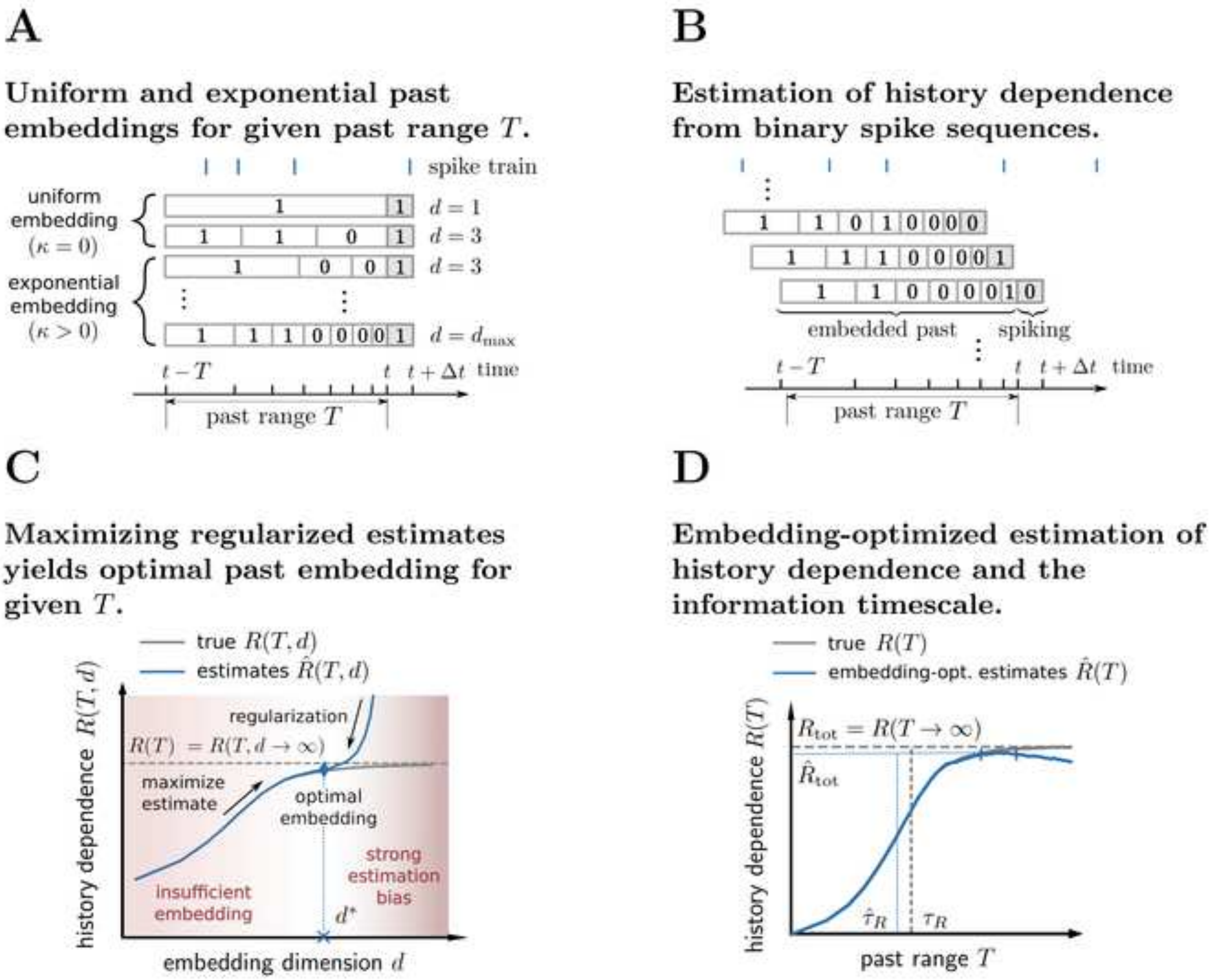
S13 Fig. Total history dependence and information timescale show no clear dependence on the firing rate, whereas the total mutual information tends to increase with the rate. Shown are the same estimates of the total history dependence R_{tot} and information timescale τ_R as in Fig 7 (Shuffling estimator with $d_{\max} = 5$) versus the firing rates of sorted units (dots). The total mutual information I_{tot} is equal to R_{tot} times the spiking entropy $H(\text{spiking})$ of the respective unit. While I_{tot} tends to increase with firing rate, no clear relation is visible for R_{tot} or τ_R. Errorbars indicate median over sorted units and 95 % bootstrapping confidence intervals on the median.

S14 Fig. Relationship between total history dependence or information timescale and standard statistical measures of neural spike trains. Estimates of the total history dependence R_{tot} tend to decrease with the median interspike interval (ISI), and to increase with the coefficient of variation C_V. This result is expected for a measure of history dependence, because a shorter median ISI indicates that spikes tend to occur together, and a higher C_V indicates a deviation from independent Poisson spiking. In contrast, the information timescale τ_R tends to increase with the autocorrelation time, as expected, with no clear relation to the median ISI or the coefficient of variation C_V. However, the correlation between the measures depends on the recorded system. For example in retina ($n = 111$), R_{tot} is significantly anti-correlated with the median ISI (Pearson correlation coefficient: $r = -0.69$, $p < 10^{-5}$) and strongly correlated with the coefficient of variation C_V ($r = 0.90$, $p < 10^{-5}$), and τ_R is significantly correlated with the autocorrelation time τ_C ($r = 0.75$, $p < 10^{-5}$). In contrast, for mouse primary visual cortex ($n = 142$), we found no significant correlations between any of these measures. Results are shown for the Shuffling estimator with $d_{\mathrm{max}} = 5$, and $T_0 = 10\,\mathrm{ms}$. Errorbars indicate median over sorted units and 95 % bootstrapping confidence intervals on the median.

S15 Fig. Excluding short-term contributions helps to differentiate the timescales for different recorded systems. By only considering gains $\Delta R(T)$ for past ranges $T > T_0$ when computing the information timescale τ_R, short-term effects that are related to the refractory period and different firing modes are excluded. The higher T_0, the higher is the distance in the median τ_R between systems (especially between salamander retina and mouse primiary visual cortex). This is because both timescales τ_R and τ_C increase with T_0 for CA1 and primary visual cortex, whereas they decrease for retina. The same holds for the autocorrelation time τ_C, where only delays $T > T_0$ were considered when fitting an exponential decay to the autocorrelograms. Note that if the decay is perfectly exponential, then T_0 does not affect the results. Estimates of R_{tot} and τ_R are shown for the Shuffling estimator with $d_{\mathrm{max}} = 5$. Errorbars indicate median over sorted units and 95 % bootstrapping confidence intervals on the median.

S16 Fig. Total history dependence decreases for small time bins Δt. The choice of the time bin Δt of the spiking activity has little effect on the information timescale τ_R, whereas the total history dependence R_{tot} decreases for small time bins $\Delta t < 5\,\mathrm{ms}$. This is consistent across experiments. The smaller the time bin, the higher the risk that noise in the spike emission reduces the overall predictability or history dependence in the spiking, whereas an overly large time bin holds the risk of destroying coding relevant time information in the spike train. Thus, we chose the smallest time bin $\Delta t = 5\,\mathrm{ms}$ that does not yet show a substantial decrease in R_{tot}. We do not plot results for higher Δt, because for higher Δt we observed many instances of multiple spikes in the same time bin. Results are shown for the Shuffling estimator with $d_{\mathrm{max}} = 5$, and $T_0 = 10\,\mathrm{ms}$. Errorbars indicate median over sorted units and 95 % bootstrapping confidence intervals on the median.


A
spike train
autocorrelation $C(T_i)$ / lagged mutual information $L(T_i)$
spiking$(t-T_i)$
1
0
spiking(t)
$t-T_i$
$t-T_{i-1}$
t
$t+\Delta t$
time
time lag T_i
B
gain
$\Delta R(T_i)$
history dependence $R(T_i)$
past(t,T_i)
0
spiking(t)
past(t,T_{i-1})
0
$t-T_i$
$t-T_{i-1}$
t
$t+\Delta t$
time
past range T_i
C
$C(T)$
0.1
0.0
τ_C
10
100
1000
$L(T)/H$
0.02
0.00
τ_L
10
100
1000
time lag T [ms]
D
$R(T)$
R_{tot}
0.1
0.0
$\Delta R(T_i)$
T_{i-1}
T_i
10
100
1000
gain $\Delta R(T)$
0.02
0.00
τ_R
10
100
1000
past range T [ms]



A

Uniform and exponential past embeddings for given past range T.

B

Estimation of history dependence from binary spike sequences.

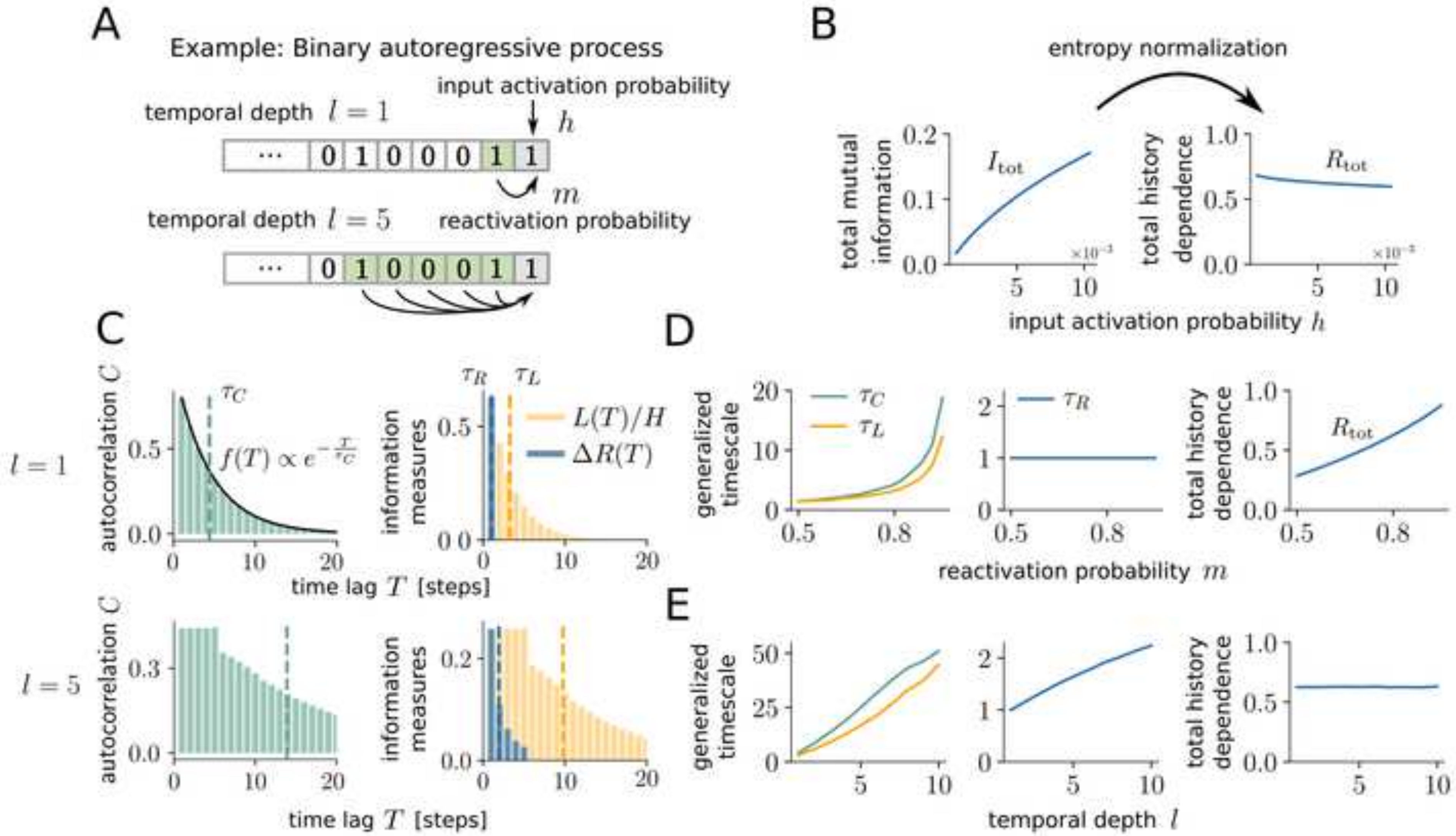
C

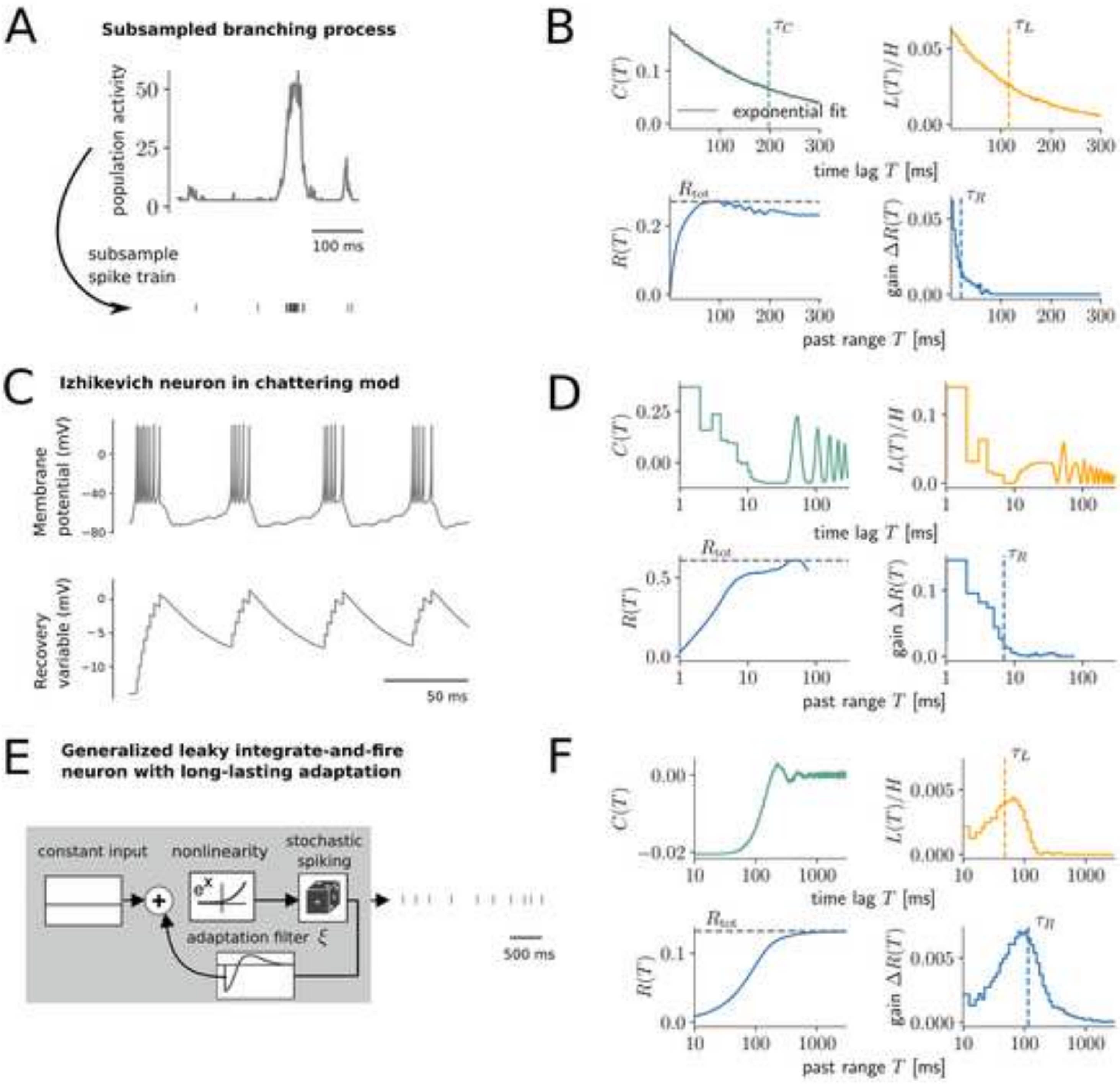
Maximizing regularized estimates yields optimal past embedding for given T.

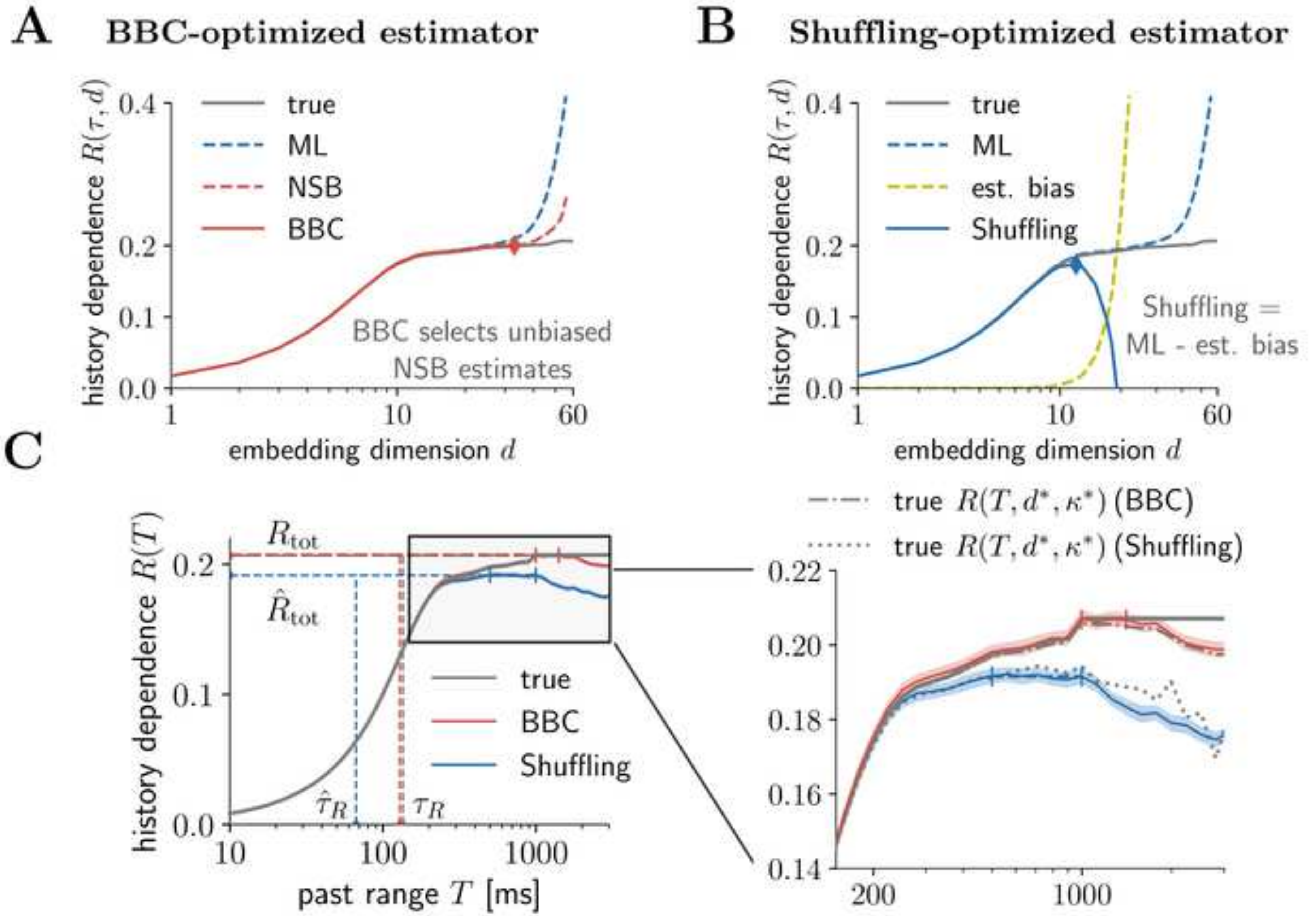
D

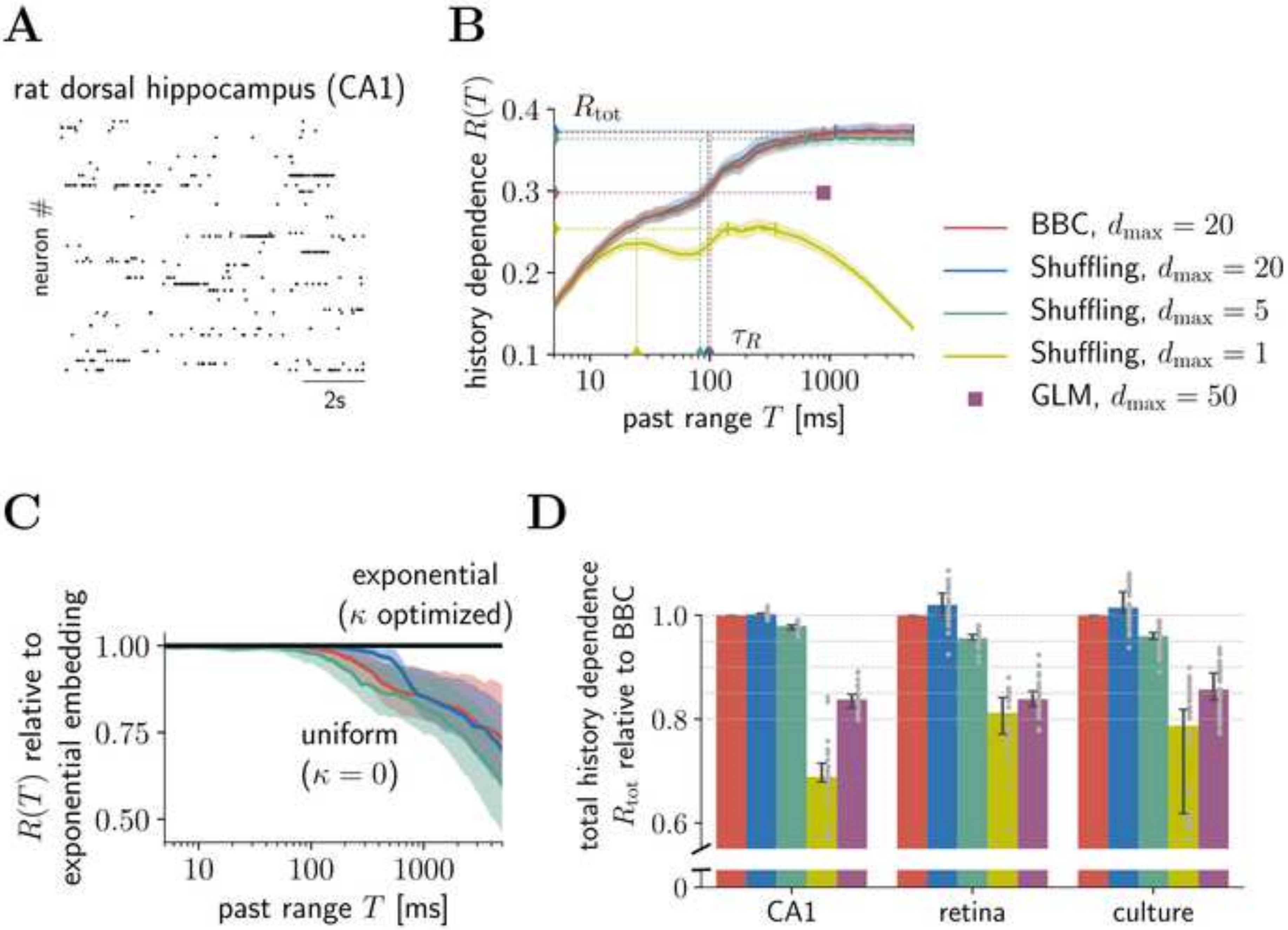
Embedding-optimized estimation of history dependence and the information timescale.

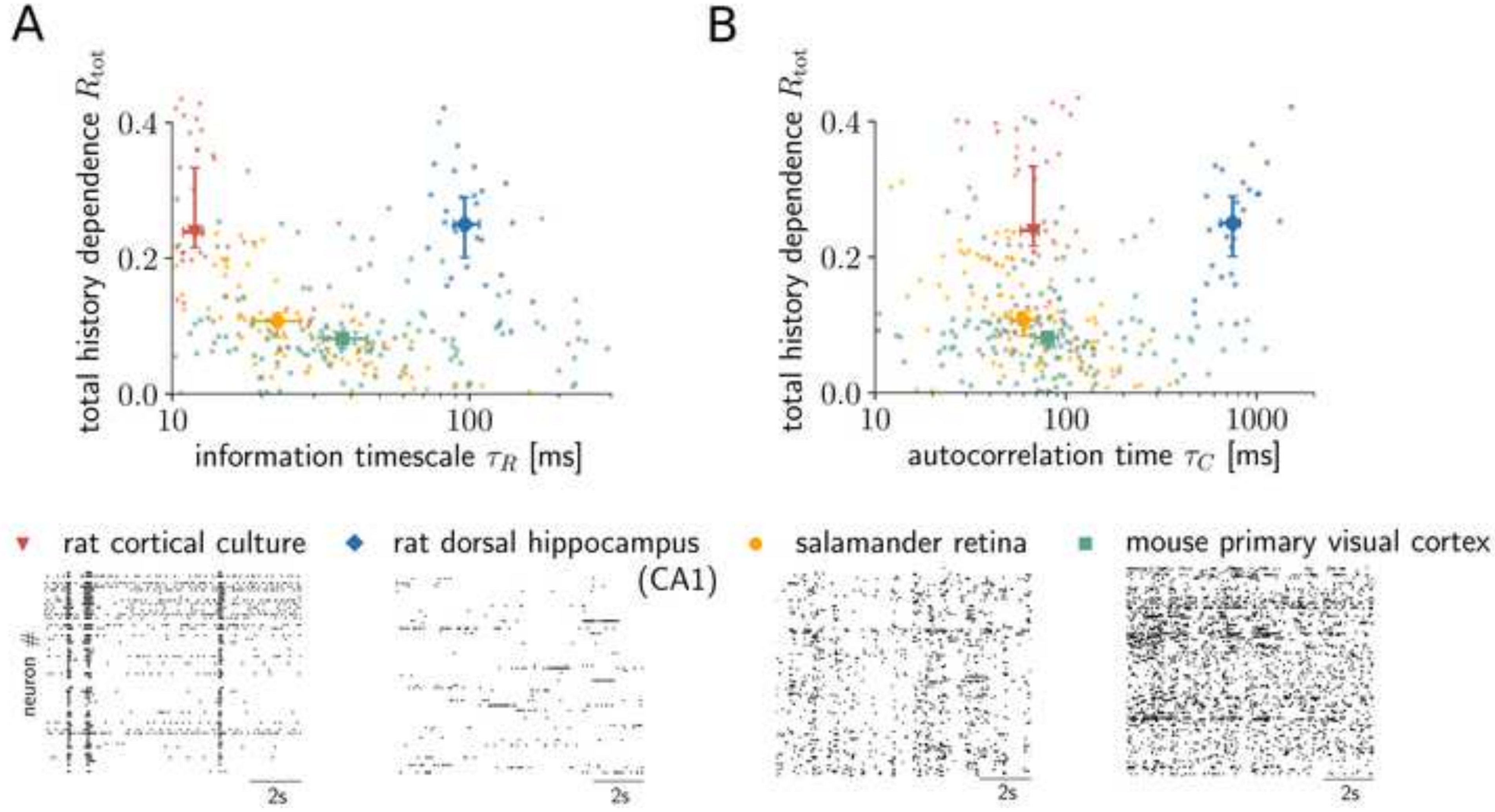
true $R(T)$
embedding-opt. estimates $\hat{R}(T)$
history dependence $R(T)$
$R_{\mathrm{tot}} = R(T \to \infty)$
$\hat{R}_{\mathrm{tot}}$
$\hat{\tau}_R$
τ_R
past range T

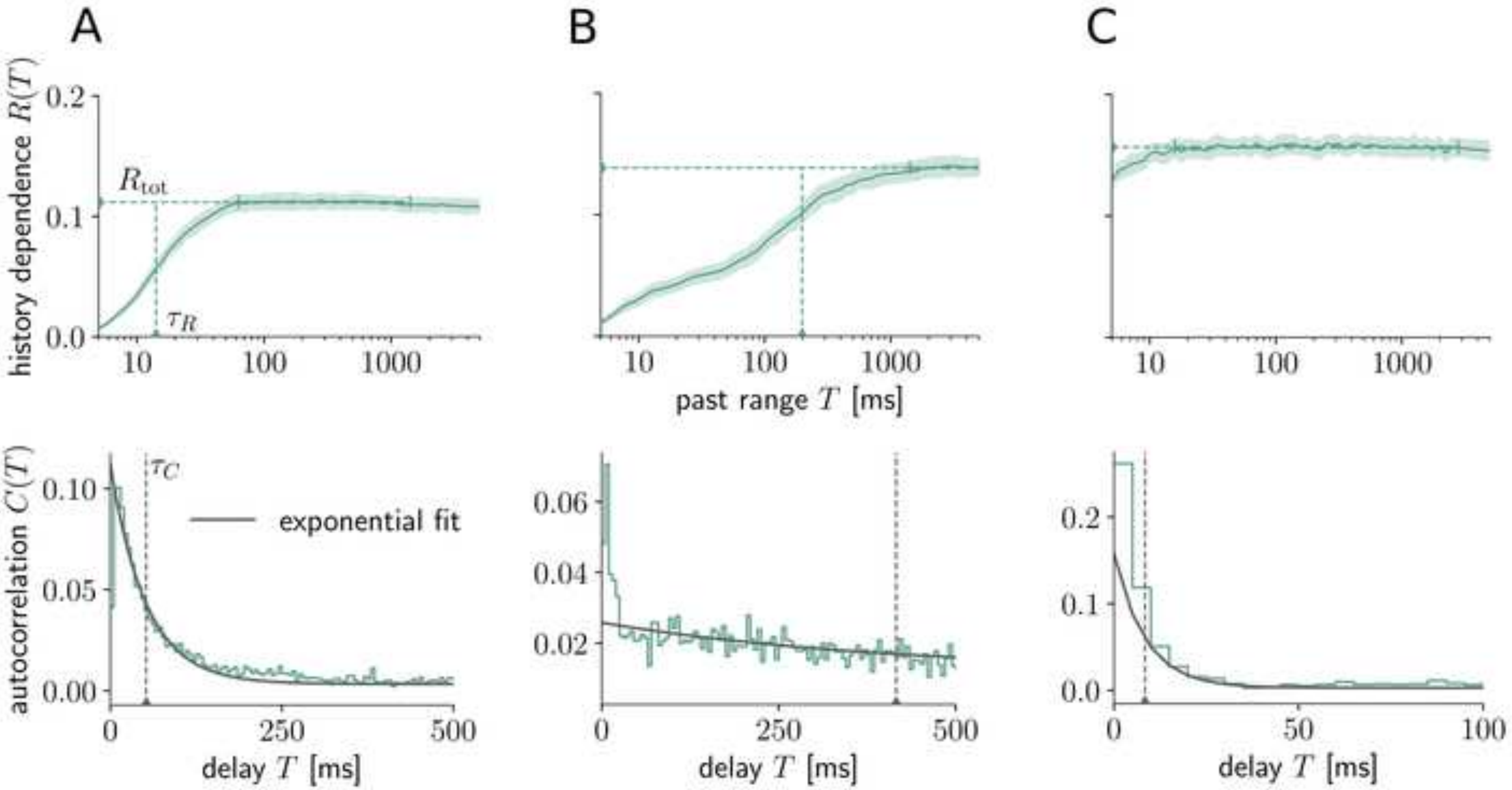

A
Example: Binary autoregressive process
input activation probability
temporal depth $l = 1$
h
··· 0 1 0 0 0 1 1
m
temporal depth $l = 5$
reactivation probability
··· 0 1 0 0 0 1 1
B
entropy normalization
total mutual information
I_{tot}
total history dependence
R_{tot}
0.2
0.1
0.0
1.0
0.5
0.0
$\times10^{-3}$
5
10
input activation probability h
C
$l = 1$
autocorrelation C
τ_C
$f(T) \propto e^{-\frac{T}{\tau_C}}$
0.5
0.0
0
10
20
information measures
τ_R
τ_L
$L(T)/H$
$\Delta R(T)$
time lag T [steps]
$l = 5$
autocorrelation C
0.3
0.0
information measures
0.2
0.0
time lag T [steps]
D
generalized timescale
τ_C
τ_L
20
10
0
0.5
0.8
τ_R
2
1
0
total history dependence
R_{tot}
1.0
0.5
0.0
reactivation probability m
E
generalized timescale
50
25
0
5
10
2
1
0
total history dependence
1.0
0.5
0.0
temporal depth l


A
Subsampled branching process
population activity
50
25
0
100 ms
subsample
spike train
B
C(T)
0.1
0.0
τ_C
exponential fit
100
200
300
L(T)/H
0.05
0.00
τ_L
time lag T [ms]
R(T)
0.2
0.0
R_{tot}
gain ΔR(T)
τ_R
past range T [ms]
C
Izhikevich neuron in chattering mod
Membrane potential (mV)
0
-40
-80
Recovery variable (mV)
0
-5
-10
50 ms
D
C(T)
0.25
0.00
L(T)/H
0.1
0.0
1
10
100
time lag T [ms]
R(T)
0.5
0.0
R_{tot}
gain ΔR(T)
0.1
0.0
τ_R
past range T [ms]
E
Generalized leaky integrate-and-fire
neuron with long-lasting adaptation
constant input
nonlinearity
stochastic spiking
e^x
adaptation filter ξ
500 ms
F
C(T)
0.00
−0.02
10
100
1000
L(T)/H
0.005
0.000
τ_L
time lag T [ms]
R(T)
0.1
0.0
R_{tot}
gain ΔR(T)
0.005
0.000
τ_R
past range T [ms]


A
BBC-optimized estimator
history dependence $R(\tau, d)$
0.4
0.2
0.1
0.0
true
ML
NSB
BBC
BBC selects unbiased
NSB estimates
1
10
60
embedding dimension d
B
Shuffling-optimized estimator
history dependence $R(\tau, d)$
0.4
0.2
0.1
0.0
true
ML
est. bias
Shuffling
Shuffling =
ML - est. bias
1
10
60
embedding dimension d
C
history dependence $R(T)$
R_{tot}
$\hat{R}_{\text{tot}}$
0.2
0.1
0.0
$\hat{\tau}_R$
τ_R
true
BBC
Shuffling
10
100
1000
past range T [ms]
true $R(T, d^*, \kappa^*)$ (BBC)
true $R(T, d^*, \kappa^*)$ (Shuffling)
0.22
0.20
0.18
0.16
0.14
200
1000


A
rat dorsal hippocampus (CA1)
neuron #
2s
B
history dependence $R(T)$
R_{tot}
τ_R
0.4
0.3
0.2
0.1
10
100
1000
past range T [ms]
BBC, $d_{\mathrm{max}} = 20$
Shuffling, $d_{\mathrm{max}} = 20$
Shuffling, $d_{\mathrm{max}} = 5$
Shuffling, $d_{\mathrm{max}} = 1$
GLM, $d_{\mathrm{max}} = 50$
C
$R(T)$ relative to exponential embedding
exponential (κ optimized)
uniform ($\kappa = 0$)
1.00
0.75
0.50
10
100
1000
past range T [ms]
D
total history dependence R_{tot} relative to BBC
1.0
0.8
0.6
0
CA1
retina
culture


A
B
total history dependence R_{tot}
0.0
0.2
0.4
10
100
information timescale τ_R [ms]
total history dependence R_{tot}
0.0
0.2
0.4
10
100
1000
autocorrelation time τ_C [ms]
▼ rat cortical culture
◆ rat dorsal hippocampus (CA1)
● salamander retina
■ mouse primary visual cortex
neuron #
2s
2s
2s
2s

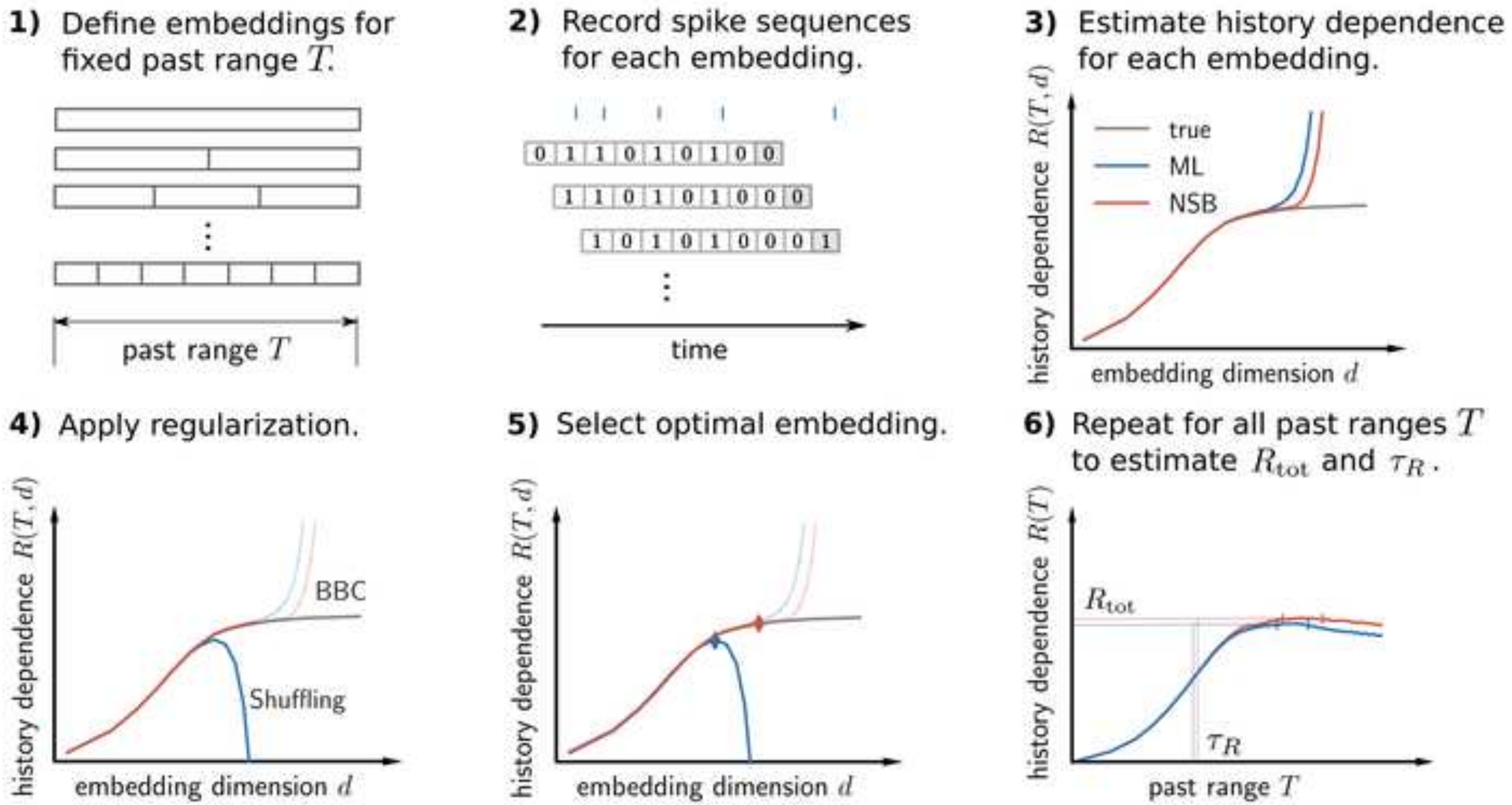

A
B
C
history dependence $R(T)$
0.2
0.1
0.0
R_{tot}
τ_R
10
100
1000
past range T [ms]
autocorrelation $C(T)$
0.10
0.05
0.00
τ_C
exponential fit
0
250
500
delay T [ms]
0.06
0.04
0.02
0.2
0.1
0.0
50
100

1) Embedding optimization: The embedding of past-spiking activity should be individually optimized to each spike train, in order to account for very different spiking statistics. This also applies to other information metrics like transfer entropy [52].

2) Regularization: Estimates have to be reliable lower bounds, otherwise one cannot interpret the results (apply Bayesian bias criterion or Shuffling correction).

3) Exponential embedding: Given the limitations on the number of bins, a non-uniform embedding is required to capture long-lasting dependencies. An exponential embedding with max. 5 bins is typically a good compromise between accuracy and computation speed, and enables embedding optimization for large, highly parallel spike recordings.

4) Data requirements: For practical purpose, spike recordings should be sufficiently long (at least 10 minutes). If several recordings are to be analyzed, these should be of similar length to allow for a meaningful comparison of history dependence and its timescale between recordings.



1) Define embeddings for fixed past range T.

2) Record spike sequences for each embedding.

0 1 1 0 1 0 1 0 0

1 1 0 1 0 1 0 0 0

1 0 1 0 1 0 0 0 1

3) Estimate history dependence for each embedding.

4) Apply regularization.

5) Select optimal embedding.

6) Repeat for all past ranges T to estimate R_{tot} and τ_R .

Click here to access/download

Supporting Information - Compressed/ZIP File Archive

Supporting information.zip

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
E-mail: lucas.rudelt@ds.mpg.de (LR), viola.priesemann@ds.mpg.de (VP)

Abstract

Information processing can leave distinct footprints on the ~~statistical history dependence in single neuronspiking. Statistical history dependence can be quantified using information theory, but its estimation from experimental recordings~~ statistics of neural spiking. For example, efficient coding minimizes the statistical dependencies on the spiking history, while temporal integration of information may require the maintenance of information over different timescales. To investigate these footprints, we developed a novel approach to quantify history dependence within the spiking of a single neuron, using the mutual information between the entire past and current spiking. This measure captures how much past information is necessary to predict current spiking. In contrast, classical time-lagged measures of temporal dependence like the autocorrelation capture how long—potentially redundant—past information can still be read out. Strikingly, we find for model neurons that our method disentangles the *strength* and *timescale* of history dependence, whereas the two are mixed in classical approaches. When applying the method to experimental data, which are necessarily of limited size, a reliable estimation of mutual information is only possible for a ~~reduced representation~~ coarse temporal binning 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~~ To still account for the vastly different spiking statistics and potentially long history dependence of living neurons, we developed an embedding-optimization approach that does not only ~~quantify non-linear and higher-order dependencies~~vary the number and size, 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~~ an exponential stretching of past bins. For 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~~ we found that the strength and timescale of history dependence indeed can vary independently across experimental preparations. While hippocampus indicated strong and long history dependence, in visual cortex it was weak and short, while in vitro the history dependence was strong but short. This work enables an information theoretic characterization of history dependence ~~, showed interesting differences between recorded systems, which points at systematic differences in information processing between these systems. We~~ in recorded spike trains, which captures a footprint of information processing that is beyond time-lagged measures of temporal dependence. To facilitate the application of the method, we provide practical guidelines ~~in this paper~~ and a toolbox~~for Python3 at 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 [1–3]. 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]~~ [5–7] and associative learning [8]. In addition to its magnitude, the ~~temporal depth~~ timescale of history dependence provides an important footprint of processing at different processing stages in the brain ~~[9]~~ [9–11]. This is because higher-level processing requires integrating information on longer timescales than lower-level processing [12]. Therefore, history dependence in neural spiking should reach further into the past for neurons involved in higher level processing ~~[9, 13]~~ [9, 13]. Quantifying history dependence and its timescale 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~~ Often, history dependence is characterized by how much spiking is correlated with spiking with a certain time lag

[14,15]. From the decay time of this lagged correlation, one obtains an intrinsic timescale of how long past information can still be read out [9–11,16]. However, to quantify not only a timescale of statistical dependence, but also its strength, one has to ~~compute~~ quantify how much of a neuron's spiking depends on its *entire past*. Here, this is done with the mutual information between the spiking of a neuron and its own past ~~[5,6,17].~~ [17], also called active information storage [5–7], or predictive information [18,19].

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 [20,21], 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 [22–26]. The downside is that this requires a lot of data, otherwise estimates can be severely biased [27,28]. A lot of work has been devoted to finding less biased estimates, either by correcting bias [28–31], or by using Bayesian inference [32–34]. 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 [35]. For example, many studies infer history dependence and transfer entropy by embedding the past spiking using a single bin [26,36].

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~~ methods summary where we introduce the measure of history dependence and the information timescale, as well as the embedding optimization 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 first compare the measure of history dependence with classical time-lagged measures of temporal dependence on different models of neural spiking activity. Second, 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~~the information timescale. For the reader interested in applying the method, we provide practical guidelines in Fig 9 and 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 [37].

Methods summary

Definition of history dependence. First, we define ~~the history dependence R~~ history dependence $R(T)$ in the spiking of a single neuron. We quantify history dependence ~~R~~ based on the mutual information ~~I(spiking; past) between the current spiking~~

$$I(\text{spiking}; \text{past}(T)) = H(\text{spiking}) - H(\text{spiking}|\text{past}(T)) \tag{1}$$

between current spiking in a time bin $[t, t+\Delta t)$ 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(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~~ in a past range $[t-T, t)$ (Fig 1B). Here, we assume stationarity and ergodicity, such that the measure is an average over all times t. This mutual information is also called active information storage [5], and is related to the predictive information [18,19]. It quantifies how much of the current spiking information H(spiking) can be predicted from past spiking. The spiking information is given by the Shannon entropy [38] ~~for current spiking reads~~

$$H(\text{spiking}) = -p(\text{spike}) \log_2 p(\text{spike}) - (1 - p(\text{spike})) \log_2 (1 - p(\text{spike})), \tag{2}$$

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(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(spiking|past) (see Materials and methods~~H(spiking|past(T)) (see Materials and methods) quantifies the average spiking information (in the sense of entropy) that ~~would be transmitted when history dependence~~ remains when dependencies on past spiking ~~is~~ are taken into account. Note that ~~history dependence~~ past dependencies can only reduce the average spiking information, i.e. ~~H(spiking|past) $\leq$ H(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~~ H(spiking|past(T)) $\leq H$(spiking). The difference between the two then gives the amount of spiking information that is redundant or entirely predictable from the past. To ~~quantify history dependence R, we chose the normalized mutual information~~ transform this measure of information into a measure of statistical dependence, ~~because it can easily be compared across recordings of neurons with very different firing rates. Moreover~~we normalize the mutual information by the entropy H(spiking) and define history dependence $R(T)$ as

$$R(T) \equiv \frac{I(\text{spiking};\text{past}(T))}{H(\text{spiking})} = 1 - \frac{H(\text{spiking}|\text{past}(T))}{H(\text{spiking})} \in [0,1]. \tag{3}$$

While the mutual information quantifies the *amount* of predictable information, $R(T)$ gives the *proportion* of spiking information that is predictable or redundant with past spiking. As such, it interpolates between the following intuitive extreme cases: ~~$R = 0$~~ $R(T) = 0$ corresponds to independent and ~~$R = 1$~~ $R(T) = 1$ to entirely predictable spiking. Moreover, while the entropy and thus the mutual information I(spiking; past(T)) increases with the firing rate (see S13 Fig. for an example on real data), the normalized $R(T)$ is comparable across recordings of neurons with very different firing rates. Finally, all the above measures can depend on the size of the time bin Δt, which discretizes the current spiking activity in time. Too small a time bin holds the risk that noise in the spike emission reduces the overall predictability or history dependence, whereas an overly large time bin holds the risk of destroying coding relevant time information in the neuron's spike train. Thus, we chose the smallest time bin $\Delta t = 5\,\text{ms}$ that does not yet show a decrease in history dependence (S16 Fig.).

~~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 10.~~

~~**Workflow of past-embedding optimization to estimate history dependence and its temporal depth.**~~

Fig 1. Illustration of history dependence and related measures in a neural spike train. (A) For the analysis, spiking is represented by 0 or 1 in a small time bin Δt (grey box). Autocorrelation $C(T)$ or the lagged mutual information $L(T)$ quantify the statistical dependence of spiking on past spiking in a single past bin with time lag T_i (green box). (B) In contrast, history dependence $R(T_i)$ quantifies the dependence of spiking on the entire spiking history in a past range T_i. The gain in history dependence $\Delta R(T_i) = R(T_i) - R(T_{i-1})$ quantifies the increase in history dependence by increasing the past range from T_{i-1} to T_i, and is defined in analogy to the lagged measures. (C) Autocorrelation $C(T)$ and lagged mutual information $L(T)$ for a typical example neuron (mouse, primary visual cortex). Both measures decay with increasing T, where $L(T)$ decays slightly faster due to the non-linearity of the mutual information. Timescales τ_C and τ_L (vertical dashed lines) can be computed either by fitting an exponential decay (autocorrelation) or by using the generalized timescale (lagged mutual information). (D) In contrast, history dependence $R(T)$ increases monotonically for systematically increasing past range T, until it saturates at the total history dependence R_{tot}. From $R(T)$, the gain $\Delta R(T_i)$ can be computed between increasing past ranges T_{i-1} and T_i (grey dashed lines). The gain $\Delta R(T)$ decays to zero like the time-lagged measures, with information timescale τ_R (dashed line).

Total history dependence and the information timescale. Here, we introduce measures to quantify the strength and the timescale of history dependence independently. First, note that the history dependence $R(T)$ monotonically increases with the past range T (Fig 1D), until it converges to the *total history dependence*

$$R_{\text{tot}} \equiv \lim_{T\to\infty} R(T). \tag{4}$$

The total history dependence R_{tot} quantifies the proportion of predictable spiking information once the entire past is taken into account.

While the history dependence $R(T)$ is monotonously increasing, the *gain* in history dependence $\Delta R(T_i) \equiv R(T_i) - R(T_{i-1})$ between two past ranges $T_i > T_{i-1}$ tends to decrease, and eventually decreases to zero for $T_i, T_{i-1} \to \infty$ (Fig 1D). This is in analogy to time-lagged measures of temporal dependence such as the autocorrelation $C(T)$ or lagged mutual information $L(T)$ (Fig 1A,C). Moreover, because $R(T)$ is monotonically increasing, the gain cannot be negative, i.e. $\Delta R(T) \geq 0$. From $\Delta R(T_i)$, we quantify a characteristic timescale τ_R of history dependence similar to an autocorrelation time. In analogy to the integrated autocorrelation time [39], we define the *generalized timescale*

$$\tau_R \equiv \sum_{i=1}^{n} \bar{T}_i \frac{\Delta R(T_i)}{\sum_{j=1}^{n} \Delta R(T_j)} - T_0. \tag{5}$$

as the average of past ranges $\bar{T}_i = (T_i + T_{i-1})/2$, weighted with their gain $\Delta R(T_i) = R(T_i) - R(T_{i-1})$. Here, steps between two past ranges T_{i-1} and T_i should be chosen small enough, and summing the middle points $\bar{T}_i$ of the steps further reduces the error of discretization. T_0 is the starting point, i.e. is the first past range for which $R(T)$ is computed, and was set to $T_0 = 10\,\text{ms}$ to exclude short-term past dependencies like refractoriness (see Materials and methods for details). Moreover, the last past range T_n has to be high enough such that $R(T_n)$ has converged, i.e. $R(T_n) = R_{\text{tot}}$. Here, we set $T_n = 5\,\text{s}$ unless stated otherwise.

To illustrate the analogy to the autocorrelation time, we note that if the gain decays exponentially, i.e. $\Delta R(T_i) \propto \exp\left(-\frac{T_i}{\tau_{\text{auto}}}\right)$ with decay constant τ_{auto}, then $\tau_R = \tau_{\text{auto}}$ for $n \to \infty$ and sufficiently small steps $T_i - T_{i-1}$. The advantage of τ_R is that it also generalizes to cases where the decay is not exponential. Furthermore, it can be applied to any other measure of temporal dependence (e.g. the lagged mutual information) as long as the sum in Eq (5) remains finite, and the coefficients are non-negative. Note that *estimates* of $\Delta R(T_i)$ can also be negative, so we included corrections to allow a sensible estimation of τ_R (Materials and methods). Finally, since τ_R quantifies the timescale over which unique predictive information is accumulated, we refer to it as the *information timescale.*

~~Discrete~~ Binary 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 [35]. 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 ~~$[t, t + \Delta t]$~~$[t, t + \Delta t)$, 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~~ with binary 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 [28, 30]. 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

Fig 2. ~~**Illustration of embedding optimization to estimate history dependence and its temporal depth.**~~ **Illustration of embedding optimization to estimate history dependence and the information timescale.** (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~~ farther 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 $\hat{\tau}_R$ and $\hat{R}_{\text{tot}}$ of the ~~temporal depth $\hat{T}_D$, as well as the~~ information timescale τ_R and total history dependence R_{tot} (vertical and horizontal dashed lines). To compute $\hat{R}_{\text{tot}}$ we average estimates $\hat{R}(T)$ in an interval $[T_D, T_{\text{max}}]$, for which estimates $\hat{R}(T)$ reach a plateau (vertical blue bars, see Materials and methods). For high past ranges T, estimates $\hat{R}(T)$ may decrease ~~,~~ because a reliable estimation requires ~~a~~ past embeddings with reduced temporal resolution.

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 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 [31] that, within leading order of the sample size, is guaranteed to provide a lower bound to the observable history dependence (see ~~Materials and methods~~ Materials and methods for details).

Together with these regularization methods, the embedding optimization approach enables complex embeddings of past activity ~~without~~ while minimizing the risk of overestimation. See Materials and methods for details on how we used embedding optimized estimates $\hat{R}(T)$ to compute estimates $\hat{R}_{\mathrm{tot}}$ and $\hat{\tau}_R$ of the total history dependence and information timescale (Fig 2, blue dashed lines).

~~**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 $R_{\mathrm{tot}} \equiv R(T \to \infty)$ 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 $\hat{R}_{\mathrm{tot}}$ 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.~~

In the first part, we demonstrate the differences between history dependence and classical measures of temporal dependence for several models of neural spiking activity. We then benchmark the estimation of history dependence using embedding optimization on a tractable neuron model with long-lasting spike adaptation. Moreover, we compare the embedding optimization approach to existing estimation methods on a variety of extra-cellular spike recordings. In the last part, we apply this to analyze history dependence for a variety of recorded systems, and compare the results to the autocorrelation and other statistical measures on the data.

Differences between history dependence and time-lagged measures of temporal dependence

The history dependence $R(T)$ quantifies how predictable neural spiking is, given activity in a certain past range T. In contrast, time-lagged measures of temporal dependence like the autocorrelation $C(T)$ [40] or lagged mutual information $L(T)$ [41, 42] quantify the dependence of spiking on activity in a single past bin with delay T (Fig 1A,C; Materials and methods). In the following, we showcase the main differences between the two approaches.

History dependence disentangles the effects of input activation, reactivation and temporal depth of a binary autoregressive process. To show the behavior of the measures in a well controlled setup, we analyzed a simple binary autoregressive process with varying temporal depth l (Fig 3). The process evolves in discrete time steps, and has an active (1) or inactive (0) state (Fig 3A). Active states are evoked either by external input with probability h, or by internal reactivations that are triggered by activity within the past l steps. Each past activation increases the reactivation probability by m, which regulates the strength of history dependence in the process. In the following, we describe how the measures behave as we vary each of the different model parameters, and then summarize the key difference between the measures.

Fig 3. History dependence disentangles the effects of input activation, reactivation and temporal depth of a binary autoregressive process. (A) In the binary autoregressive process, the state of the next time step (grey box) is active (one) either because of an input activation with probability h, or because of an internal reactivation. The internal activation is triggered by activity in the past l time steps (green), where each active state increases the activation probability by m. (B) Increasing the input activation probability h increases the total mutual information, although input activations are random and therefore not predictable. Normalizing the total mutual information by the entropy yields the total history dependence, which decreases mildly with h. (C) Autocorrelation $C(T)$, lagged mutual information $L(T)$ and gain in history dependence $\Delta R(T)$ decay differently with the delay T. For $l = 1$ and $m = 0.8$ (top), autocorrelation $C(T)$ decays exponentially with autocorrelation time τ_C, whereas $L(T)$ decays faster due to the non-linearity of the mutual information. $\Delta R(T)$ is non-zero only for delays shorter or equal to the temporal depth of the process, with much shorter timescale τ_R. For $l = 5$, $C(T)$ and $L(T)$ plateau over the temporal depth, and then decay much slower than for $l = 1$. Again, $\Delta R(T)$ is non-zero only within the temporal depth of the process. Parameters m and h were adapted to match the firing rate and total history dependence between $l = 1$ and $l = 5$. (D) When increasing the reactivation probability m for $l = 1$, timescales of time-lagged measures τ_C and τ_L increase. For history dependence, the information timescale τ_R remains constant, but the total history R_{tot} increases. (E) When varying the temporal depth l, all timescales increased. Parameters h and m were adapted to hold the firing rate and R_{tot} constant.

The input strength h increases the firing rate and thus the spiking entropy H(spiking). This leads to a strong increase in the total mutual information

$I_{\mathrm{tot}} \equiv \lim_{T\to\infty} I(\mathrm{spiking};\mathrm{past}(T))$, whereas the total history dependence R_{tot} is normalized by the entropy and does slightly decrease (Fig 3B). This slight decrease is expected from a sensible measure of history dependence, because the input is random and has no temporal dependence. In addition, input activations may fall together with internal activations, which slightly reduces the total history dependence. In contrast, the total history dependence R_{tot} increases with the reactivation probability m, as expected (Fig 3D). For the autocorrelation, the reactivation probability m not only influences the magnitude of the correlation coefficients, but also the decay of the coefficients. For autoregressive processes (and $l = 1$), autocorrelation coefficients $C(T)$ decay exponentially [14] (Fig 3C), where the autocorrelation time $\tau_C = -\Delta t/\log(m)$ increases with m and diverges as $m \to 1$ (Fig 3D). The lagged mutual information $L(T)$ is a non-linear measure of time-lagged dependence, and has a very similar behavior as the autocorrelation, with a slightly faster decay and thus smaller generalized timescale τ_L (Fig 3C,D). Note that we normalized $L(T)$ by the spiking entropy H to make it directly comparable to $\Delta R(T)$. In contrast to the time-lagged measures, the gain in history dependence $\Delta R(T)$ is only non-zero for T smaller or equal to the true temporal depth l of the process (Fig 3C). As a consequence, the information timescale τ_R does not increase with m for fixed l (Fig 3D).

Finally, the temporal depth l controls how far into the past activations depend on their preceding activity. Indeed, we find that the information timescale τ_R increases with l as expected (Fig 3C,E). Similarly, the timescales of the time-lagged measures τ_C and τ_L increase with the temporal depth l. Note that parameters m and h were adapted for each l to keep the firing rate and total history dependence R_{tot} constant, such that differences in the timescale can be unambiguously attributed to the increase in l.

To conclude, history dependence disentangles the effects of input activation, reactivation and temporal depth, which provides a comprehensive characterization of past dependencies in the autoregressive model. This is different from the total mutual information, which lacks the entropy normalization and is sensitive to the firing rate. This is also different from time-lagged measures, whose timescales are sensitive to both, the reactivation probability m *and* the temporal depth l. The confusion of effects in the timescales is rooted in the time-lagged nature of the measures—by quantifying past dependencies out of context, $C(T)$ and $L(T)$ also capture *indirect, redundant* dependencies onto past events. Indirect, redundant dependencies arise from unique dependencies, because past states that are uniquely predictive of future activities were in turn uniquely dependent on their own past. The stronger the unique dependence, the longer the indirect dependencies reach into the past, which increases the timescale of time-lagged measures. In contrast, indirect dependencies do not contribute to the history dependence, because they add no predictive information once more-recent past is taken into account.

History dependence dismisses redundant past dependencies and captures synergistic effects. A key property of history dependence is that it evaluates past dependencies in the light of more recent past. This allows the measure to dismiss indirect, redundant past dependencies and to capture synergistic effects. In three common models of neural spiking activity, we demonstrate how this leads to a substantially different characterization of past dependencies compared to time-lagged measures of temporal dependence.

First, we simulated a subsampled branching process [14], which is a minimal model for activity propagation in neural networks and captures key properties of spiking dynamics in cortex [15]. Similar to the binary autoregressive process, active neurons activate neurons in the next time step with probability m, the so called

Fig 4. History dependence dismisses redundant past dependencies and captures synergistic effects (A,B) Analysis of a subsampled branching process. (A) The population activity was simulated as a branching process ($m = 0.98$) and subsampled to yield the spike train of a single neuron (Materials and methods). (B) Autocorrelation $C(T)$ and lagged mutual information $L(T)$ include redundant dependencies and decay much slower than the gain $\Delta R(T)$, with much longer timescales (vertical dashed lines). (C,D) Analysis of an Izhikevich neuron in chattering mode with constant input and small voltage fluctuations. The neuron fires in regular bursts of activity. (D) Time-lagged measures $C(T)$ and $L(T)$ measure both, intra- ($T < 10\,\mathrm{ms}$) and inter-burst ($T > 10\,\mathrm{ms}$) dependencies, which decay very slowly due to regularity of the firing. The gain $\Delta R(T)$ reflects that most spiking can already be predicted from intra-burst dependencies, whereas inter-burst dependencies are highly redundant. In this case, only $\Delta R(T)$ yields a sensible time scale (blue dashed line). (E,F) Analysis of a generalized leaky integrate and fire neuron with long-lasting adaptation filter ξ [3,43] and constant input. Figure adapted from [44]. (F) Here, $\Delta R(T)$ decays slower to zero than the autocorrelation $C(T)$, and is higher than $L(T)$ for long delays T. Therefore, the dependence on past spikes is stronger when taking more recent past spikes into account ($\Delta R(T)$), as when considering them independently ($L(T)$). Due to these synergistic past dependencies, $\Delta R(T)$ is the only measure that captures the long-range nature of the spike adaptation.

branching parameter, and are activated externally with some probability h. The process was simulated in time steps of $\Delta t = 4\,\mathrm{ms}$ with a population activity of 500 Hz, which was subsampled to obtain a single spike train with a firing rate of 5 Hz (Fig 4A). Similar to the binary autoregressive process, the autocorrelation decays exponentially with autocorrelation time $\tau_C = -\Delta t/\log(m) = 198\,\mathrm{ms}$, and the lagged mutual information decays slightly faster (Fig 4B). In comparison, the gain in history dependence ΔR decays much faster. When increasing the branching parameter m (for fixed firing rate), the total history dependence increased, as in the autoregressive process (S11 Fig.). Strikingly, the timescale τ_R remained constant or even decreased for larger $m > 0.967$ and thus higher autocorrelation time $\tau_C > 120\mathrm{ms}$ (S11 Fig.), which is different from the binary autoregressive process. The reason is that the branching process evolves at the population level, whereas history dependence is quantified at the single neuron level. Thereby, history dependence also captures indirect dependencies, because the own spiking history reflects the population activity. The higher the branching parameter m, the more informative past spikes are about the population activity, and the shorter is the timescale τ_R over which all the relevant information about the population activity can be collected. Thus, for the branching process, the total history dependence R_{tot} captures the influence of the branching parameter, whereas the information timescale τ_R behaves very differently from the timescales of time-lagged measures.

Second, we demonstrate the difference of history dependence to time-lagged measures on an Izhikevich neuron, which is a flexible model that can produce different neural firing patterns similar to those observed for real neurons [45]. Here, parameters were chosen according to the ”chattering mode” [45], with constant input and small voltage fluctuations (Materials and methods). The neuron fires in regular bursts of activity, with consistent timing between spikes within and between bursts (Fig 4C). While time-lagged measures capture all the regularities in spiking and oscillate with the bursts of activity, history dependence correctly captures that spiking can almost be entirely predicted from intra-burst dependencies alone (Fig 4D). History dependence dismisses the redundant inter-burst dependencies and thereby yields a sensible measure of a timescale (blue dashed line).

Finally, we analyzed a generalized leaky integrate-and-fire neuron with long-range spike adaptation (22 seconds) (Fig 4E), which reproduces spike-frequency adaptation as observed for real layer 2/3 pyramidal neurons [3,43]. For this model, time-lagged measures $C(T)$ and $L(T)$ actually decay to zero much faster than the gain in history dependence $\Delta R(T)$, which is the only measure that captures the long-range adaptation effects of the model (Fig 4F). This shows that past dependencies in this model include synergistic effects, where the dependence is stronger in the context of more recent spikes. This is most likely due to the non-linearity of the model, where past spikes cause a different adaptation when taken together as when considered as the sum of their contributions.

Thus, due to its ability to dismiss redundant past dependencies and to capture synergistic effects, history dependence really provides a complementary characterization of past dependencies compared to time-lagged measures. Importantly, because the approach better disentangles the effects of timescale and total history dependence, the results remain interpretable for very different models, and provide a more comprehensive view on past dependencies.

Embedding optimization ~~can capture long-lasting~~ captures history dependence for a ~~benchmark spiking~~ neuron model with long-lasting spike adaptation

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~~ Readers that are familiar with the bias problem of mutual information estimation might want to jump to the next part, where we validate that embedding-optimized estimates indeed capture the model's ~~history dependence for hundreds of milliseconds~~true history dependence, while being robust to systematic overestimation. As a model we chose a generalized leaky integrate-and-fire ~~neuron~~ (GLIF) model with spike frequency adaptation, whose parameters were fitted to experimental data [3,43]. The ~~neuron was driven with a constant input current to achieve an average firing rate of 4 Hz. The model neuron~~ model was chosen, because it is equipped with a long-lasting spike adaptation mechanism~~that lasts over 20 seconds, and the ground truth of the~~, and its 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)~~Materials and methods). For demonstration, we show results on a variant of the model where adaptation reaches one second into the past, and show results on the original model with a 22 second kernel in S1, S2 and S5 Figs. For simulation, the neuron was driven with a constant input current to achieve an average firing rate of 4 Hz. In the following, estimates $\hat{R}(T)$ are shown for a simulated recording of 90 minutes, whereas ~~GLM estimates~~ the true values $R(T)$ were computed on a 900 minute recording (Materials and methods).

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 $\tau = 20\,\mathrm{ms}$ (i.e. $\kappa = 0$ and $T = d \cdot \tau$). We compared estimates $\hat{R}(\tau, d)$ obtained by maximum likelihood (ML) estimation [28], or Bayesian estimation using the NSB estimator [33], with the model's true $R(\tau, d)$ ~~.~~

(Fig 5A). Both estimators accurately estimate $R(\tau, d)$ for up to ~~$d \approx 15$~~ $d \approx 20$ 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.

Fig 5. ~~**Embedding optimization accurately estimates history dependence for a generalized leaky integrate-and-fire neuron with long-lasting spike frequency adaptation [3,43].**~~ **Embedding optimization captures history dependence for a neuron model with long-lasting spike adaptation.** Results are shown for a generalized leaky integrate-and-fire (GLIF) model with long-lasting spike frequency adaptation [3,43] with a temporal depth of one second (Methods and material). (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) [28] and Bayesian (NSB) [33] 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, which are very small and thus hardly visible. (C) Embedding optimized BBC estimates $\hat{R}(T)$ (red line) yield accurate estimates of the model neuron's true history dependence $R(T)$~~for hundreds of milliseconds~~, total history dependence R_{tot} and information timescale τ_R (horizontal and vertical dashed lines). The zoom-in (right panel) shows robustness of both regularization methods: For all T the model neuron's ~~$R(T)$~~ $R(T, d^*, \kappa^*)$ lies within errorbars (BBC), or consistently above the Shuffling estimator that provides a lower bound. Here, the model's ~~$R(T)$~~ $R(T, d^*, \kappa^*)$ was computed for the optimized embedding parameters ~~$d^*, \kappa*$~~ d^*, κ^* that were selected via BBC or Shuffling, respectively (dashed lines). Shaded areas indicate ~~95 % confidence intervals~~ ± two standard deviations obtained by bootstrapping, and colored ~~dashed lines~~ vertical bars indicate past ranges over which estimates $\hat{R}(T)$ were averaged to compute $\hat{R}_{\text{tot}}$ (Materials and methods).

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 5A, red diamond), which is formalized by the Bayesian bias criterion (BBC) (~~Materials and methods~~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 [31]. The Shuffling estimator is based on the ML estimator and applies a bias correction term (Fig 5B). 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 5B, 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~~ We find that embedding-optimized BBC estimates $\hat{R}(T)$ agree well with the true $R(T)$, such that 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~~ total history dependence R_{tot} and information timescale τ_R are well estimated (Fig 5C, vertical and horizontal dashed lines). In contrast, the Shuffling estimator ~~. Notably, for both estimators, estimates $\hat{R}(T)$ agree with the true history dependence for up to several hundred milliseconds (Fig 5C). When comparing the two regularization methods (BBC and Shuffling), the BBC approach captures more history dependence.~~

~~For both regularization methods the~~ underestimates the true $R(T)$ for past ranges $T > 200\,\text{ms}$, such that the model's R_{tot} and τ_R are underestimated (blue dashed lines). For large past ranges $T > 1000\,\text{ms}$, estimates $\hat{R}(T)$ ~~decrease for high T. This is because little~~ of both estimators decrease again, because no 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 5C). 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.~~ of the embedding.

Embedding-optimized estimates ~~do not overestimate history dependence~~ are robust to overestimation 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 (Fig 5A,B).

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~~ grey dotted line, Fig 5C, 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 (S1 and S4 Figs). 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.~~e.g. by optimizing embedding parameters on ~~one~~ the first half, and computing estimates on the ~~other~~ second half of the data (S1 Fig). 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 (S1 Fig). 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~~no cross-validation was applied on the experimental data.

Estimates of ~~temporal depth~~ the information timescale are sensitive to the recording length. Finally, we also tested the impact of the recording length on ~~the value of the estimated~~ estimates $\hat{R}_{\mathrm{tot}}$ of the total history dependence ~~$\hat{R}_{\mathrm{tot}}$,~~ as well as ~~the temporal depth $\hat{T}_D$~~estimates $\hat{\tau}_R$ of the information timescale. While on recordings of 3 minutes embedding optimization still estimated $\approx 95\,\%$ of ~~$\hat{R}_{\mathrm{tot}}$ that was estimated for 90 minutes, the estimated $\hat{T}_D$ was only half of the temporal depth that was estimated for 90 minutes~~ the true R_{tot}, estimates $\hat{\tau}_R$ were only $\approx 75\,\%$ of the true τ_R (S2 Fig). ~~The temporal depth decreases for shorter recordings, because the variance of estimates increases, such that estimates $\hat{R}(T)$ saturate within errorbars for smaller~~ Thus, estimates of the information timescale τ_R are more sensitive to the recording length, because they depend on the small additional contributions to $R(T)$ for high past ranges T~~. We therefore advice to compare history dependence, and especially $\hat{T}_D$, for~~ , which are hard to estimate for short recordings. Therefore, we advice to analyze recordings of similar ~~recording length~~ length to make results on τ_R comparable across experiments. 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~~true history dependence. 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~~ the information timescale in experimental spike recordings.

Embedding optimization ~~reveals~~ is key to estimate long-lasting history dependence in extra-cellular spike recordings~~of spiking neurons~~

Here, we apply embedding optimization to long spike recordings ($\approx$ 90 minutes) from rat dorsal hippocampus layer CA1 [46, 47], salamander retina [48, 49] and in vitro recordings of rat cortical culture [50]. 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.~~

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 ~~$\hat{R}(T)$~~ $R(T)$ as a function of the past range T (see Fig 6B for an example ~~neuron~~ single unit from hippocampus layer CA1, and S6, S7 and S8 Figs for all analyzed ~~neurons~~sorted units). 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~~ $R_{\mathrm{tot}} \approx 40\%$ with a temporal depth of almost a second, and an information timescale of $\tau_R \approx 100\,\mathrm{ms}$ (Fig 6B). 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 6B,D). To compare how well the different estimation approaches could capture the total history dependence, we plotted for each ~~neuron~~ so the different estimates ~~$\hat{R}_{\mathrm{tot}}$~~ of R_{tot} relative to the corresponding BBC estimate (Fig 6D). 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~~ that was estimated using BBC (Fig 6B, 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~~Materials and methods). Since ~~embedding and model parameters for the GLM~~ parameters were optimized, we argue that the GLM underestimates history dependence because of its ~~model assumption of no interdependencies~~ specific model assumptions, i.e. no interactions between past spikes. ~~Considering that~~ Moreover, we found that the GLM performs worse than embedding optimization with only five past bins~~estimates much higher history dependence~~. Therefore, we conclude that ~~interdependencies between past events~~ for typical experimental spike trains, interactions between past spikes are important, but do not require very high temporal resolution. In the remainder of this paper we use the reduced approach (Shuffling $d_{\max} = 5$) to compare history dependence among different recorded systems.

Increasing bin sizes exponentially is crucial to estimate long-lasting history dependence~~for high past ranges~~. To demonstrate this, we plotted embedding-optimized BBC estimates ~~$\hat{R}(T)$~~ of $R(T)$ using a uniform embedding, i.e.

Fig 6. ~~**Embedding optimization reveals strong and long-lasting history dependence in experimental spike recordings.**~~ **Embedding optimization is key to estimate long-lasting history dependence in extra-cellular spike recordings.** (A) Example of recorded spiking activity in rat dorsal hippocampus layer CA1. (B) ~~Estimated~~ Estimates of history dependence ~~$\hat{R}(T)$~~ $R(T)$ for various estimators, as well as estimates of the ~~estimated~~ total history dependence ~~$\hat{R}_{\mathrm{tot}}$~~ R_{tot} and ~~temporal depth $\hat{T}_D$~~ information timescale τ_R (dashed lines) (example ~~neuron~~ single unit 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~~ indicate ± two standard ~~deviation~~ deviations obtained by bootstrapping, and ~~colored dashed lines~~ small vertical bars indicate past ranges over which estimates ~~$\hat{R}(T)$~~ of $R(T)$ were averaged to ~~compute $\hat{R}_{\mathrm{tot}}$~~estimate R_{tot} (Materials and methods). (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 ~~$\hat{R}(T)$~~ of $R(T)$ with uniform embeddings, relative to estimates obtained by optimizing exponential embeddings, for BBC with $d_{\max} = 20$ (red) and Shuffling with $d_{\max} = 20$ (blue) and $d_{\max} = 5$ (green). Shaded areas show 95 % percentiles. Median and percentiles were computed over analyzed ~~neurons~~ sorted units in CA1 ($n = 28$). (D) Comparison of ~~estimated~~ total history dependence ~~$\hat{R}_{\mathrm{tot}}$~~ R_{tot} for different estimation and embedding techniques for three different experimental recordings. For each ~~neuron~~ sorted unit (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~~ sorted units (CA1: $n = 28$; retina: $n = 111$; culture: $n = 48$).

equal bin sizes, relative to estimates obtained with exponential embedding (Fig 6C), both for BBC with $d_{\max} = 20$ (red) and Shuffling with $d_{\max} = 20$ (blue) or $d_{\max} = 5$ (green). For past ranges $T > 100$ 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 $T = 1$ s, and only around 60 % for ~~$T = 10$ s~~ $T = 5$ s (median over analyzed ~~neurons~~ sorted units in CA1). Therefore, we argue that an exponential embedding is crucial ~~when assessing the temporal depth of history dependence in neural spiking activity.~~ for estimating long-lasting history dependence

~~**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 dependenceand temporal depth for spike recordings from different species and brain areas.~~

~~Embedding optimization reveals clear differences in~~ Together, total history dependence and its timescale show clear differences between recorded systems and individual ~~neurons~~sorted units

Finally, we present results from diverse ~~electrophysiological~~ extracellular spike recordings that show interesting differences in history dependence between ~~neurons~~ sorted units of different recorded systems. In addition to recordings from rat dorsal hippocampus layer CA1, salamander retina and rat cortical culture, we analyzed ~~neural spike trains recorded in~~ sorted units in a recording of mouse primary visual cortex using the novel ~~Neuropixel~~ Neuropixels probe [51]. 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~~ information timescale (Fig 7~~).~~ ~~Neurons~~ A). Sorted units in cortical culture and hippocampus layer CA1 have high total history dependence R_{tot} with median over ~~neurons~~ sorted units of $\approx 24\,\%$ and $\approx 25\,\%$, whereas ~~neurons~~ sorted units in retina and primary visual cortex have typically low R_{tot} of $\approx 11\,\%$ and $\approx 8\,\%$. In terms of ~~temporal depth, neurons~~ the information timescale τ_R, sorted units in hippocampus layer CA1 display much higher ~~temporal depth T_D~~ τ_R with a median of ~~$\approx 450\,\text{ms}$ than neurons~~ $\approx 96\,\text{ms}$ than units in cortical culture with median ~~temporal depth of $\approx 60\,\text{ms}$~~τ_R of $\approx 12\,\text{ms}$. Similarly, ~~neurons~~ sorted units in primary visual cortex have higher ~~T_D~~ τ_R with median of ~~$\approx 160\,\text{ms}$ than neurons~~ $\approx 37\,\text{ms}$ than units in retina with median of ~~$\approx 70\,\text{ms}$~~$\approx 23\,\text{ms}$. 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~~ total history dependence and the information timescale varied independently among recorded systems, and studying them in isolation would miss differences between recorded systems, whereas considering them jointly allows to distinguish the different systems~~in terms of history dependence~~. Moreover, no clear differentiation between cortical culture, retina and primary visual cortex is possible using the autocorrelation time τ_C (Fig 7B), with medians $\tau_C \approx 68\,\text{ms}$ (culture), $\tau_C \approx 60\,\text{ms}$ (retina) and $\tau_C \approx 80\,\text{ms}$ (primary visual cortex), respectively.

To better understand how other well-established statistical measures relate to the total history dependence R_{tot} and the information timescale τ_R, we show R_{tot} and τ_R versus the median interspike inteval (ISI), the coefficient of variation $C_V = \sigma_{\text{ISI}}/\mu_{\text{ISI}}$ of the ISI distribution, and the autocorrelation time τ_C in S14 Fig.. Estimates of the total history dependence R_{tot} tend to decrease with the median ISI, and to increase with the coefficient of variation C_V. This result is expected for a measure of history dependence, because a shorter median ISI indicates that spikes tend to occur together, and a higher C_V indicates a deviation from independent Poisson spiking. In contrast, the information timescale τ_R tends to increase with the autocorrelation time, as expected, with no clear relation to the median ISI or the coefficient of variation C_V. However, the correlation between the measures depends on the recorded system. For example in retina ($n = 111$), R_{tot} is significantly anti-correlated with the median ISI (Pearson correlation coefficient: $r = -0.69$, $p < 10^{-5}$) and strongly correlated with the

Fig 7. ~~**Total history dependence and temporal depth show clear differences between recorded systems.**~~ **Together, total history dependence and its timescale show clear differences between recorded systems.** (A) Embedding-optimized Shuffling estimates ($d_{\max} = 5$) of the total history dependence R_{tot} are plotted against the ~~temporal depth~~ information timescale τ_R for individual ~~neurons~~ sorted units (dots) from four different recorded systems (raster plots show ~~spiketrains~~ spike trains of ~~the recorded neurons~~different sorted units). No clear relationship between the two quantities is visible. The analysis shows systematic differences between the recorded systems: ~~Neurons~~ sorted units in rat cortical culture ($n = 48$) and rat dorsal hippocampus layer CA1 ($n = 28$) have higher median total history dependence than ~~neurons~~ units in salamander retina ($n = 111$) and mouse primary visual cortex ($n = 142$). At the same time, ~~neurons~~ sorted units in cortical culture and retina show smaller ~~temporal depth~~ timescale than ~~neurons~~ units in primary visual cortex, and much smaller ~~temporal depth~~ timescale than ~~neurons~~ units in hippocampus layer CA1. Overall, recorded systems are clearly distinguishable when jointly considering the total history dependence and ~~temporal depth~~information timescale. ~~Error bars~~ (B) Total history dependence R_{tot} versus the autocorrelation time τ_C shows no clear relation between the two quantities, similar to the information timescale τ_R. Also, the autocorrelation time gives the same relation in timescale between retina, primary visual cortex and CA1, whereas the cortical culture has a higher timescale (different order of medians on the x-axis). In general, recorded systems are harder to differentiate in terms of the autocorrelation time τ_C as compared to τ_R. Errorbars indicate median over ~~neurons~~ sorted units and 95 % bootstrapping confidence intervals on the median.

coefficient of variation C_V ($r = 0.90$, $p < 10^{-5}$), and τ_R is significantly correlated with the autocorrelation time τ_C ($r = 0.75$, $p < 10^{-5}$). In contrast, for mouse primary visual cortex ($n = 142$), we found no significant correlations between any of these measures. Thus, the relation between R_{tot} or τ_R and the established measures is not systematic, and therefore one cannot replace the history dependence by any of them.

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~~ sorted units in primary visual cortex (Fig 8, see S9 Fig for all analyzed ~~neurons~~ sorted units in primary visual cortex). In particular, ~~neurons~~ sorted units display different signatures of history dependence $R(T)$ as a function of the past range T. For some ~~neurons~~units, history dependence builds up on short past ranges T (e.g. Fig 8A), for some it only shows for higher T (e.g. Fig 8B), and for some it already saturates for very short T (e.g. Fig 8C). A similar behavior is captured by the autocorrelation $C(T)$ (Fig 8, second row). The rapid saturation in Fig 8C indicates history dependence due to bursty firing, which can also be seen by strong positive correlation with past spikes for short delays T (Fig 8C, bottom). To exclude the effects of different firing modes or refractoryness on the information timescale, we only considered past ranges $T > T_0 = 10\,\mathrm{ms}$ when estimating τ_R, or delays $T > T_0 = 10\,\mathrm{ms}$ when fitting an exponential decay to $C(T)$ to estimate τ_C. The reason is that differences in the integration of past information are expected to show for larger T. This agrees with the observation that timescales among recorded systems were much more similar if one instead sets $T_0 = 0\,\mathrm{ms}$, whereas they showed clear differences for $T_0 = 10\,\mathrm{ms}$ or $T_0 = 20\,\mathrm{ms}$ (S15 Fig.).

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 $T \approx 200$ ms in ~~EC~~ CA1 (Fig 6B), possibly reflecting phase information in the theta cycles [46, 47]. Thus, the analysis does not only serve to investigate

Fig 8. ~~**Distinct signatures of history dependence for different neurons within mouse primary visual cortex.**~~ **Distinct signatures of history dependence for different sorted units within mouse primary visual cortex.** (Top) Embedding-optimized estimates of $R(T)$ reveal distinct signatures of history dependence for three different ~~neurons~~ sorted units (A,B,C) within a single recorded system (mouse primary visual cortex). In particular, ~~neurons~~ sorted units have similar total history dependence ~~$\hat{R}_{\mathrm{tot}}$~~R_{tot}, but differ vastly in the ~~estimated temporal depth $\hat{T}_D$~~information timescale τ_R (horizontal and vertical dashed lines). Note that for unit C, τ_R is smaller than 5 ms and thus doesn't appear in the plot. Shaded areas indicate ± two standard ~~deviation~~ deviations obtained by bootstrapping, and vertical bars indicate the ~~dashed line indicates past ranges~~ interval over which estimates ~~$\hat{R}(T)$~~ of $R(T)$ were averaged to ~~compute $\hat{R}_{\mathrm{tot}}$~~estimate R_{tot} (Materials and methods). Estimates were computed with the Shuffling estimator and $d_{\max} = 5$. (Bottom) Autocorrelograms for the same sorted units (A,B, and C, respectively) roughly show an exponential decay, which was fitted (solid grey line) to estimate the autocorrelation time τ_C (grey dashed line). Similar to the information timescale τ_R, only coefficients for delays larger than $T_0 = 10$ ms were considered during fitting.

differences in history dependence between recorded systems, but also resolves clear differences between sorted units. This could be used to investigate differences in information processing between different cortical layers, different neuron types or neurons with different receptive field properties.

Overall, ~~this demonstrates~~ our results demonstrate that embedding optimization is powerful enough to reveal clear differences in history dependence between ~~neurons~~ sorted units of different recorded systems, but also between ~~neurons~~ units within the same system. Even more importantly, because ~~neurons~~ units 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~~experimental data, we developed a method where the embedding of past spiking is optimized for each individual ~~neuron~~spike train. 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.

Based on our results, we arrived at practical guidelines that are summarized in Fig 9. In the following, we contrast history dependence ~~R with pairwise~~ $R(T)$ with time-lagged measures such as the ~~auto-correlation~~autocorrelation in more detail, clearly discussing the advantages—but also the limitations of the approach. We then discuss how one can relate estimated history dependence to neural coding and information processing ~~at~~ based on the example data sets analyzed in this paper.

~~**Why quantify history dependence and not the auto-correlation or auto-information?**~~ ~~First,~~

Advantages and limitations of history dependence in comparison to the autocorrelation and lagged mutual information. A key difference between

1) Embedding optimization: The embedding of past-spiking activity should be individually optimized to each spike train, in order to account for very different spiking statistics. This also applies to other information metrics like transfer entropy [52].

2) Regularization: Estimates have to be reliable lower bounds, otherwise one cannot interpret the results (apply Bayesian bias criterion or Shuffling correction).

3) Exponential embedding: Given the limitations on the number of bins, a non-uniform embedding is required to capture long-lasting dependencies. An exponential embedding with max. 5 bins is typically a good compromise between accuracy and computation speed, and enables embedding optimization for large, highly parallel spike recordings.

4) Data requirements: For practical purpose, spike recordings should be sufficiently long (at least 10 minutes). If several recordings are to be analyzed, these should be of similar length to allow for a meaningful comparison of history dependence and its timescale between recordings.

Fig 9. Practical guidelines for the estimation of history dependence in single neuron spiking activity. More details regarding the individual points can be found at the end of Materials and methods.

history dependence $R(T)$ ~~captures an important footprint of neural coding that is not captured by pairwise dependency measures such as the auto-correlation [40] or the auto- or delayed mutual information [41]. History dependence~~ and the autocorrelation or lagged mutual information is that $R(T)$ quantifies statistical dependencies between current spiking and ~~past spiking in the entire~~ the *entire past spiking* in a past range T ~~.~~ ~~From~~ (Fig 1B). This has the following advantages as a measure of statistical dependence, and as a footprint of information processing in single neuron spiking. First, $R(T)$ allows to compute the total history dependence, which, from a coding perspective, ~~it gives~~ represents the redundancy of neural spiking with all past spikes~~in the past range T,~~ ~~;~~ or how much ~~past information in T is integrated~~ of the past information is also represented when emitting a spike. Second, because past spikes are considered jointly, $R(T)$ captures synergistic effects and dismisses redundant past information (Fig 4). Finally, we found that this enables $R(T)$ to disentangle the strength and timescale of history dependence for the binary autoregressive process. (Fig 3). In contrast, ~~auto-correlation or auto-information~~ autocorrelation $C(T)$ or lagged mutual information $L(T)$ quantify the statistical ~~dependency~~ dependence of neural spiking ~~onto~~ on 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~~ with delay T, without considering any of the other bins (Fig 1A). Thereby, they miss synergistic effects; and they quantify redundant past dependencies that vanish once spiking activity in more recent past is taken into ~~consideration.~~ account (Fig 4). As a consequence, the timescales of these measures reflect both, the strength and the temporal depth of history dependence in the binary autoregressive process (Fig 3).

~~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~~ Moreover, technically, the autocorrelation time ~~[9,53], 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 [14,15,54]. For any deviations from the simple model , however, the autocorrelation might~~ τ_C depends on fitting exponential decay to coefficients $C(T)$. Computing the autocorrelation time with the generalized timescale is difficult, because coefficients $C(T)$ can be negative, and are too noisy for large delays T. While model fitting is in general more data efficient than the model-free estimation presented here, it can also produce biased and unreliable estimates [16]. Furthermore, when the coefficients do 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 more complex model has to be fitted [53], or the analysis simply cannot be applied. In contrast, the generalized timescale can be directly applied to estimates of the history dependence $R(T)$ to yield the information timescale τ_R without any further assumptions or fitting models. However, we found that estimates of τ_R can depend strongly on the estimation method and embedding dimension (S12 Fig.) and the size of the data set (S2 and S3 Figs). The dependence on data size is not so strong for the practical approach of optimizing up to $d_{\max} = 5$ past bins, but still we recommend to use data sets of similar length when aiming for comparability across experiments. Moreover, there might be cases where a model-free estimation of the true timescale might be infeasible because of the complexity of past dependencies (S2 Fig, neuron with a 22 seconds past kernel). In this case, only $\approx 80\,\%$ of the true timescale could be estimated on a 90 minute recording.

~~A~~ Another downside of quantifying the history dependence ~~R~~ $R(T)$ is that its estimation requires more data than fitting the autocorrelation time τ_C. To make best use of the limited data, we here devised ~~an~~ the embedding optimization approach that allows to find ~~an~~ the most efficient representation of past spiking for the estimation of history dependence. ~~Nonetheless~~Even so, we found empirically that a minimum of 10 minutes of recorded spiking activity are advisable to ~~allow~~ achieve a meaningful quantification of history dependence and its ~~temporal depth ()~~ timescale (S2 and S3 Figs). In addition, for shorter recordings, the analysis can lead to mild overestimation due to over-optimizing embedding parameters on noisy estimates (S2 Fig). This overestimation can, however, be avoided by cross-validation, which we find to be particularly relevant for the Bayesian bias criterion (BBC) estimator. Finally, our approach uses an embedding model that ranges from uniform embedding to an embedding with exponentially stretching past bins—assuming that past information farther into the past requires less temporal resolution. This embedding model might be inappropriate if for example spiking depends on the exact timing of distant past spikes, with gaps in time where past spikes are irrelevant. In such a case, embedding optimization could be used to optimize more complex embedding models that can also account for this kind of spiking statistics.

Differences in total history dependence and ~~temporal depth~~ information timescale between data sets agree with ideas from neural coding and hierarchical information processing. First, we found that the ~~estimated~~ total history dependence ~~$\hat{R}_{\text{tot}}$~~ R_{tot} clearly differs among the experimental data sets. Notably, ~~$\hat{R}_{\text{tot}}$~~ R_{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 [1,55] and neural adaptation for temporal whitening as observed in experiments [3,56]. In contrast, ~~$\hat{R}_{\text{tot}}$~~ R_{tot} was high for neurons in dorsal hippocampus (layer CA1) and cortical culture. In CA1, the original study [47] 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 ~~$\hat{R}_{\text{tot}}$~~ R_{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 [57], or to recognize statistical structure for associative learning [8]. In cortical culture, neurons are exclusively driven by recurrent input and exhibit strong bursts in the population activity [58]. 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 ~~$\hat{R}_{\text{tot}}$~~ R_{tot} does indeed provide a footprint of neural coding and information processing.

Second, we observed that the ~~temporal depth T_D of history dependence~~ information timescale τ_R increases from retina (~~$\approx$ 70 ms~~$\approx$ 23 ms) to primary visual cortex (~~$\approx$ 160 ms) to EC ($\approx$ 450 ms~~$\approx$ 37 ms) to CA1 ($\approx$ 96 ms), in agreement with the idea of a temporal hierarchy in neural information processing [12]. These results qualitatively agree with similar results obtained for the autocorrelation time of spontaneous activity [9], although the information timescales are overall much smaller than the autocorrelation times. Our results ~~indicate~~ suggest that the hierarchy of intrinsic timescales ~~is also reflected~~ could also show 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~~ its strength, as well as ~~the temporal depth~~its timescale. 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~~ Neuropixels probe [59], or even across cortical layers within a single area. To this end we provide a toolbox for Python3 [37]~~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~~the information timescale. 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~~ $\Delta R(T_i) \equiv R(T_i) - R(T_{i-1})$: gain in history dependence
- τ_R: information timescale or generalized timescale of history dependence $R(T)$
- $L(T) \equiv I(X; X_{-T})$: lagged mutual information with time lag T ~~for which $R(T) = R_{\text{tot}}$~~
- τ_L: generalized timescale of lagged mutual information $L(T)$

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 ~~$T \in [\hat{T}_D, T_{\text{max}}]$~~ $T \in [T_D, T_{\text{max}}]$, with interval of saturated estimates $[T_D, T_{\text{max}}]$
- $\hat{\tau}_R$: estimated information timescale

Basic definitions

Definition of history dependence. We ~~estimate~~ quantify 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.

$$R(T) \equiv \frac{I(X, \boldsymbol{X}^{-T})}{H(X)} = 1 - \frac{H(X|\boldsymbol{X}^{-T})}{H(X)}. \tag{6}$$

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 [21, 60], i.e.

$$\begin{aligned} I(X, \boldsymbol{X}^{-T}) &= H(X) - H(X|\boldsymbol{X}^{-T}) \qquad (7)\\ &= \sum_{x\in\{0,1\}} p(x) \log_2 \frac{1}{p(x)} - \sum_{\boldsymbol{x}^{-T}\in\{0,1\}^d} p(x, \boldsymbol{x}^{-T}) \log_2 \frac{1}{p(x|\boldsymbol{x}^{-T})} \qquad (8)\\ &= \sum_{x\in\{0,1\}} \sum_{\boldsymbol{x}^{-T}\in\{0,1\}^d} p(x, \boldsymbol{x}^{-T}) \log_2 \frac{p(x|\boldsymbol{x}^{-T})}{p(x)} \qquad (9)\\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \log_2 \frac{p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T})}{p(x_{t_n})}. \qquad (10) \end{aligned}$$

Here, $x_{t_n} \in \{0, 1\}$ indicates the presence of a spike in a small interval ~~$[t_n, t_n + \Delta t]$~~ $[t_n, t_n + \Delta t)$ with $\Delta t = 5\,\mathrm{ms}$ 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.

Definition of lagged mutual information. The lagged mutual information $L(T)$ [41] for a stationary neural spike trains is defined as the mutual information between present spiking X and past spiking X_{-T} with delay T, i.e. =0

$$\begin{aligned} L(T) &\equiv I(X; X_{-T}) \qquad (11)\\ &= \sum_{x\in\{0,1\}} \sum_{x_{-T}\in\{0,1\}} p(x, x_{-T}) \log_2 \frac{p(x|x_{-T})}{p(x)} \qquad (12)\\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \log_2 \frac{p(x_{t_n}|x_{t_n-T})}{p(x_{t_n})}. \qquad (13) \end{aligned}$$

Here, $x_{t_n} \in \{0, 1\}$ indicates the presence of a spike in a time bin $[t_n, t_n + \Delta t)$ and $x_{t_n-T} \in \{0, 1\}$ the presence of a spike in a single past bin $[t_n - T, t_n - T + \Delta t)$ at times $t_n = n\Delta t$ that are shifted by Δt. In analogy to $R(T)$, one can apply the generalized timescale to the lagged mutual information to obtain a timescale τ_L with

$$\tau_L \equiv \sum_{i=1}^{n} \bar{T}_i \frac{L(T_i)}{\sum_{i=j}^{n} L(T_j)} - T_0. \tag{14}$$

Definition of autocorrelation. The autocorrelation $C(T)$ for a stationary neural spike trains is defined as

$$C(T) = \frac{\mathrm{Cov}[x_{t_n}, x_{t_n-T}]}{\mathrm{Var}[x_{t_n}]} = \frac{\langle x_{t_n} x_{t_n-T}\rangle - \langle x_{t_n}\rangle^2}{\langle x_{t_n}^2\rangle - \langle x_{t_n}\rangle^2} \tag{15}$$

with delay T and x_{t_n} and x_{t_n-T} as above. For an exponentially decaying autocorrelation $C(T) \propto \exp\left(-\frac{T}{\tau_C}\right)$, τ_C is called *autocorrelation time.*

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

$$\Gamma_T(\theta) : \mathcal{F}_T \to S^d \tag{16}$$

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 ~~$\tau_j = \tau 10^{(j-1)\kappa}$~~$\tau_j = \tau_1 10^{(j-1)\kappa}$. Here, ~~$\tau$~~ τ_1 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.

$$p(x_t|\boldsymbol{x}_{t,\theta}^{-T}, \boldsymbol{x}_{t,\nu}^{-T}) = p(x_t|\boldsymbol{x}_{t,\theta}^{-T}) \tag{17}$$

for any other past embedding $\boldsymbol{x}_{t,\nu}^{-T}$. If Eq (17) 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 [61], i.e.

$$D_{KL}\left[p(x_t|\boldsymbol{x}_t^{-T})||p(x_t|\boldsymbol{x}_{t,\theta}^{-T})\right] = \sum_{x_t \in \{0,1\}} p(x_t|\boldsymbol{x}_t^{-T}) \log_2 \frac{p(x_t|\boldsymbol{x}_t^{-T})}{p(x_t|\boldsymbol{x}_{t,\theta}^{-T})} \geq 0, \tag{18}$$

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

$$\begin{aligned}
0 \leq & \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{x_{t_n}\in\{0,1\}} p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T}) \log_2 \frac{p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T})}{p(x_{t_n}|\boldsymbol{x}_{t_n,\theta}^{-T})} \quad (19)\\
= & \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{x_{t_n}\in\{0,1\}} p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T}, \boldsymbol{x}_{t_n,\theta}^{-T}) \log_2 \frac{1}{p(x_{t_n}|\boldsymbol{x}_{t_n,\theta}^{-T})} \quad (20)\\
& - \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{x_{t_n}\in\{0,1\}} p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T}) \log_2 \frac{1}{p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T})} \quad (21)\\
= & H(X|\boldsymbol{X}_\theta^{-T}) - H(X|\boldsymbol{X}^{-T}), \quad (22)
\end{aligned}$$

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.

$$R(T,\theta) \equiv 1 - \frac{H(X|\boldsymbol{X}_\theta^{-T})}{H(X)} \leq 1 - \frac{H(X|\boldsymbol{X}^{-T})}{H(X)} = R(T). \tag{23}$$

~~**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~~

$$R_{\mathrm{tot}} \equiv \lim_{T\to\infty} R(T),$$

~~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.~~

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

~~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

$$\theta^* = \arg\max_{\theta} \hat{R}(T,\theta). \tag{24}$$

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 (S1 Fig). 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 and the information timescale. ~~Using the embedding-optimized~~ When estimating history dependence $R(T)$ from data, there are some adjustments required to estimate the total history dependence R_{tot} and the information timescale τ_R.

First, estimates $\hat{R}(T)$ are not guaranteed to converge for large past ranges T, but might decrease due to a reduced resolution of embeddings for higher T (Fig 2D). Thus, we estimated an interval $[T_D, T_{\max}]$ for which estimates have converged. Here, the temporal depth T_D ~~is estimated as the minimum past range~~ and the upper bound $T_{\max}$ are the first and the last past ranges T for which estimates $\hat{R}(T)$ ~~lies~~ are within one standard deviation of the ~~maximum estimated history dependence~~highest estimate $\hat{R}_{\max}$, i.e.

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

~~with~~

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

$\hat{R}(T) \geq \hat{R}_{\max} - \sigma_{\hat{R}_{\max}}$ (Fig 2D, vertical blue bars). 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~~ Bootstrap confidence intervals). From this interval, an estimate of the total history dependence ~~was estimated~~ $\hat{R}_{\text{tot}}$ is obtained 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~~ $T \in [T_D, T_{\max}]$ (Fig 2D, vertical dashed blue line).

Second, noisy estimates $\hat{R}(T)$ ~~lies within standard error of the maximum estimated history dependence, i. e.~~

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

~~This avoids averaging over estimates that are systematically underestimated because of limited resolution for high past ranges~~are not guaranteed to be monotonously increasing, such that increments $\Delta\hat{R}(T)$ can be negative. Moreover, noisy estimates can lead to positive $\Delta\hat{R}(T)$ even though the true $R(T)$ has already converged to R_{tot}. This can have a huge effect on the estimated information timescale $\hat{\tau}_R$ if one simply uses these estimates in Eq (5). To avoid this, we use knowledge about the behavior of the true $R(T)$ when estimating $\Delta R(T)$. In particular, we set estimates $\hat{R}(T)$ equal to the largest previous estimate $\hat{R}(T')$ for $T' < T$ if they fall below it, and equal to $\hat{R}_{\mathrm{tot}}$ if they are larger than $\hat{R}_{\mathrm{tot}}$. This enforces that the estimated gain $\Delta\hat{R}(T) \geq 0$ is non-negative, and excludes spurious gain for high T due to noisy estimates.

Finally, the information timescale τ_R can crucially depend on the choice of the minimum past range T_0 in the sum in Eq (5). A $T_0 > 0$ larger than zero allows to ignore short term effects on the history dependence such as the refractory period or different firing modes, which we found beneficial for resolving differences in the timescale among different recorded systems (S15 Fig.). In contrast, if the decay is truly exponential, than τ_R is independent of T_0. In this paper, we chose $T_0 = 10\,\mathrm{ms}$ to exclude short term effects, while also not excluding too much past information.

Workflow. ~~The estimation workflow using embedding optimization can be summarized by the following sequence of steps (Fig 10):~~

~~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 $x_{t_n}, \boldsymbol{x}^{-T}_{t_n,\theta}$ 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 $\hat{R}_{\mathrm{tot}}$.~~

The estimation workflow using embedding optimization is summarized in (Fig 10).

Fig 10. Workflow of past-embedding optimization to estimate history dependence and its temporal depth. 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 $x_{t_n}, \boldsymbol{x}^{-T}_{t_n,\theta}$ for all time steps t_n in the recording. **3)** Use the frequencies of the recorded sequences to estimate history dependence for each embedding, either using maximum likelihood (ML), or fully Bayesian estimation (NSB). **4)** Apply regularization, i.e. the Bayesian bias criterion (BBC) or Shuffling bias correction, 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 of $R(T)$. **6)** Repeat the estimation for a set of past ranges T to compute estimates of the information timescale τ_R and the total history dependence R_{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 $H(X|\boldsymbol{X}^{-T}_{\theta})$ from data. The estimation of the binary entropy only requires the average firing probability $p(x{=}1) = r\Delta t$ with

$$\hat{H}(X) = -r\Delta t \log_2 r\Delta t - (1 - r\Delta t)\log_2(1 - r\Delta t), \tag{25}$$

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

$$H(X|\boldsymbol{X}_\theta^{-T}) = H(X, \boldsymbol{X}_\theta^{-T}) - H(\boldsymbol{X}_\theta^{-T}) \tag{26}$$

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~~ patterns $a_k \in \{0,1\}^{d+1}$. The entropy $H(X, \boldsymbol{X}_\theta^{-T})$ then reads

$$H(X, \boldsymbol{X}_\theta^{-T}) = H(\boldsymbol{\pi}) = -\sum_{k=1}^{K} \pi_k \log_2 \pi_k. \tag{27}$$

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 $[t_n, t_n + \Delta t]$~~ time step $[t_n, t_n + \Delta t)$ 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

$$p(\boldsymbol{n}|\boldsymbol{\pi}) \propto \prod_{k=1}^{K} \pi_k^{n_k}. \tag{28}$$

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

$$\hat{H}_{\text{ML}}(X, \boldsymbol{X}_\theta^{-T}) = -\sum_{k=1}^{K} \frac{n_k}{N} \log_2 \frac{n_k}{N} \tag{29}$$

or history dependence

$$\hat{R}_{\text{ML}}(T, \theta) = 1 - \frac{\hat{H}_{\text{ML}}(X, \boldsymbol{X}_\theta^{-T}) - \hat{H}_{\text{ML}}(\boldsymbol{X}_\theta^{-T})}{\hat{H}(X)}. \tag{30}$$

The ML estimator has the right asymptotic properties [28, 62], but is known to underestimate the entropy severely when data is limited [28, 63]. 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

$$\text{bias}\text{Bias}\left[\hat{H}_{\mathrm{ML}}(X, \boldsymbol{X}_{\theta}^{-T})\right] \leq 0. \tag{31}$$

The negative bias in the entropy, which is largest for the highest-dimensional joint entropy $\hat{H}_{\mathrm{ML}}(X, \boldsymbol{X}_{\theta}^{-T})$, then typically leads to severe overestimation of the mutual information and history dependence [27, 64]. 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)

$$\hat{H}_{\mathrm{MMSE}}(\boldsymbol{n}) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{n}) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) \frac{p(\boldsymbol{n}|\boldsymbol{\pi})p(\boldsymbol{\pi})}{\int d\boldsymbol{\pi}' p(\boldsymbol{n}|\boldsymbol{\pi}')p(\boldsymbol{\pi}')}. \tag{32}$$

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 $\boldsymbol{n}$

$$p(\boldsymbol{\pi}|\boldsymbol{n}) = \frac{p(\boldsymbol{n}|\boldsymbol{\pi})p(\boldsymbol{\pi})}{\int d\boldsymbol{\pi}' p(\boldsymbol{n}|\boldsymbol{\pi}')p(\boldsymbol{\pi}')}. \tag{33}$$

The probability for i.i.d. observations $\boldsymbol{n}$ from an underlying distribution $\boldsymbol{\pi}$ is given by the multinomial distribution in Eq (28).

If the prior $p(\boldsymbol{\pi})$ is a conjugate prior to the multinomial likelihood, then the high dimensional integral of Eq (32) can be evaluated analytically [32]. This is true for a class of priors called Dirichlet priors, and in particular for symmetric Dirichlet priors

$$p(\boldsymbol{\pi}|\beta) \propto \prod_{k=1}^{K} \pi_k^{\beta-1}. \tag{34}$$

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 [65]. 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

$$\xi(\beta) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) p(\boldsymbol{\pi}|\beta) = \psi_0(K\beta + 1) - \psi_0(\beta + 1), \tag{35}$$

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

$$p_{NSB}(\boldsymbol{\pi}) \propto \int d\beta \left|\frac{\partial \xi}{\partial \beta}\right| p(\boldsymbol{\pi}|\beta), \tag{36}$$

$$\frac{\partial \xi}{\partial \beta} = K\psi_1(K\beta + 1) - \psi_1(\beta + 1), \tag{37}$$

that weights Dirichlet priors to be flat with respect to the expected entropy $\xi(\beta)$. Since the variance of this expectation vanishes for $K \gg 1$ [65], 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

$$\hat{H}_{NSB}(\boldsymbol{n}) = \int d\boldsymbol{\pi} H(\boldsymbol{\pi}) \frac{p(\boldsymbol{n}|\boldsymbol{\pi}) p_{NSB}(\boldsymbol{\pi})}{\int d\boldsymbol{\pi}' p(\boldsymbol{n}|\boldsymbol{\pi}') p_{NSB}(\boldsymbol{\pi}')} \tag{38}$$

$$= \frac{\int d\beta \frac{d\xi}{d\beta}(\beta) \hat{H}(\beta) \rho(\beta, \boldsymbol{n})}{\int d\beta' \frac{d\xi}{d\beta}(\beta') \rho(\beta', \boldsymbol{n})}. \tag{39}$$

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

$$\rho(\beta, \boldsymbol{n}) := \frac{\Gamma(K\beta)}{\Gamma(N + K\beta)} \prod_{i=1}^{K} \frac{\Gamma(n_i + \beta)}{\Gamma(\beta)} \tag{40}$$

$$\propto \int d\boldsymbol{\pi}\, p(\boldsymbol{n}|\boldsymbol{\pi})\, p(\boldsymbol{\pi}|\beta) = p(\boldsymbol{n}|\beta), \tag{41}$$

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

$$\hat{H}(\beta) = \sum_{i=1}^{K} \frac{n_i + \beta}{N + K\beta} [\psi_0(N + K\beta + 1) - \psi_0(n_i + \beta + 1)]. \tag{42}$$

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

$$\hat{R}_{\mathrm{NSB}}(T, \theta) = 1 - \frac{\hat{H}_{\mathrm{NSB}}(X, \boldsymbol{X}_{\theta}^{-T}) - \hat{H}_{\mathrm{NSB}}(\boldsymbol{X}_{\theta}^{-T})}{\hat{H}(X)}. \tag{43}$$

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 [37]. 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 [34].

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

$$|\hat{R}_{\mathrm{NSB}}(T, \theta) - \hat{R}_{\mathrm{ML}}(T, \theta)| > p \cdot \hat{R}_{\mathrm{NSB}}(T, \theta). \tag{44}$$

Similarly, we define the BBC estimator

$$\hat{R}_{\mathrm{BBC}}(T, \theta) \equiv \begin{cases} \hat{R}_{\mathrm{NSB}}(T, \theta) & \text{if} \quad \hat{R}_{\mathrm{NSB}}(T, \theta) - \hat{R}_{\mathrm{ML}}(T, \theta) \leq p \cdot \hat{R}_{\mathrm{NSB}}(T, \theta), \\ 0 \quad \text{otherwise.} \end{cases} \tag{45}$$

This estimator is designed to be unbiased, and can thus can be used for embedding optimization in Eq (24). 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 5 and S2 Fig).

Shuffling estimator. The Shuffling estimator was originally proposed in [31] 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

$$\hat{R}_{\mathrm{ML}}(T, \theta) = \frac{1}{\hat{H}(X)} \left(\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}) - \hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X) \right) \tag{46}$$

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 [28, 63], we know that

$$\begin{aligned} \mathrm{Bias}[\hat{R}_{\mathrm{ML}}(T, \theta)] &= \mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T})] - \mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)] && (47) \\ &\leq -\mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)]. && (48) \end{aligned}$$

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 [31] 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

$$p_{\mathrm{sh}}(\boldsymbol{x}_\theta^{-T}|x) = \prod_{j=1}^{d} p(x_{\theta,j}^{-T}|x), \tag{49}$$

and the corresponding entropy is

$$H(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X) = \sum_{j=1}^{d} H(X_{j\theta,j}{}^{-T}|X). \tag{50}$$

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

$$\Delta\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X)] \equiv \hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X) - H(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X) \tag{51}$$

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 [27, 64]

$$\mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)] = -\frac{1}{2N\ln 2}\sum_{x\in\{0,1\}}\left(\tilde{K}(x)-1\right) + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{52}$$

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

$$\mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X)] \lesssim \mathrm{Bias}[\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X)]. \tag{53}$$

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

$$\hat{H}_{\mathrm{ML,sh}}(\boldsymbol{X}_\theta^{-T}|X) \equiv \hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}|X) - \Delta\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_{\theta,\mathrm{sh}}^{-T}|X), \tag{54}$$

and use this to estimate history dependence

$$\hat{R}_{\mathrm{Shuffling}}(T,\theta) \equiv \frac{1}{\hat{H}(X)}\left(\hat{H}_{\mathrm{ML}}(\boldsymbol{X}_\theta^{-T}) - \hat{H}_{\mathrm{ML,sh}}(\boldsymbol{X}_\theta^{-T}|X)\right). \tag{55}$$

Because of Eq (48) and Eq (53), we know that this estimator is negatively biased in leading order

$$\hat{R}_{\mathrm{Shuffling}}(T,\theta) \lesssim 0 \tag{56}$$

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 [21]. 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

$$\hat{H}(X|\boldsymbol{X}^{-T}, \nu) = \frac{1}{N} \sum_{n=1}^{N} \log_2 p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T}, \nu)^{-1} \tag{57}$$

which one can plug into Eq (6) to obtain an estimate of the history dependence. The strong law of large numbers [60] 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

$$\nu^* = \arg\max_{\nu} \sum_{n=1}^{N} \log_2 p(x_{t_n}|\boldsymbol{x}_{t_n}^{-T}, \nu). \tag{58}$$

We here consider a generalized linear model (GLM) with exponential link function that has successfully been applied to make predictions in neural spiking data [20] and can be used for the estimation of directed, causal information [21]. In a GLM with past dependencies, the spiking probability at time t is described by the instantaneous rate or conditional intensity function

$$\lambda(t|\boldsymbol{x}_t^{-T}, \nu) = \lim_{\delta t \to 0} \frac{p(\hat{t} \in [t, t+\delta t]|\boldsymbol{x}_t^{-T}, \nu)}{\delta t}. \tag{59}$$

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

$$p(x_t{=}1|\boldsymbol{x}_t^{-T}, \nu) = \frac{\lambda(t|\boldsymbol{x}_t^{-T}, \nu)\Delta t}{1 + \lambda(t|\boldsymbol{x}_t^{-T}, \nu)\Delta t}. \tag{60}$$

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

$$\lambda(t|\boldsymbol{x}_t^{-T}, \mu, \boldsymbol{h}) = \exp\left(\mu + \sum_{j=1}^{d} h_j x_{t,j}^{-T}\right), \tag{61}$$

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 [66]. In particular, for given past range T, we choose embedding parameters d^*, κ^* that minimize

$$\mathrm{BIC}(d, \kappa) = (d+1)\log_2 N - 2\mathcal{L}^*(d, \kappa), \tag{62}$$

where N is the number of samples and

$$\mathcal{L}^*(d,\kappa) = \sum_{n=1}^{N} \log_2 p(x_{t_n}|\boldsymbol{x}^{-T}_{t_n,d,\kappa}, \mu^*, \boldsymbol{h}^*) \tag{63}$$

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

$$\hat{H}_{\mathrm{GLM}}(X|\boldsymbol{X}^{-T}_{d^*,\kappa^*}) = -\frac{1}{N}\mathcal{L}^*(d^*,\kappa^*), \tag{64}$$

which results in the GLM estimator of history dependence

$$\hat{R}_{\mathrm{GLM}}(T) = 1 - \frac{\hat{H}_{\mathrm{GLM}}(X|\boldsymbol{X}^{-T}_{d^*,\kappa^*})}{\hat{H}(X)}. \tag{65}$$

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 [67]. To obtain bootstrap samples, we first compute all the binary sequences $(x_{t_n}, \boldsymbol{x}^{-T}_{t_n,\theta})$ 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~~ interspike 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 (S10 Fig). 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 (??)~~past range T with the highest estimated history dependence $R(T)$. 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 (S10 Fig). 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 (S1 Fig). To avoid this type of overestimation, we apply one round of 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 $(x_{t_n}, \boldsymbol{x}^{-T}_{t_n,\theta})$ 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 (S1 Fig).

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 [43].

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

$$C\dot{V} = -g_L(V - V_R) + I_{\mathrm{ext}}(t). \tag{66}$$

Here, V_R denotes the resting potential and $I_{\mathrm{ext}}(t)$ 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~~ interspike intervals in the simulation were all larger than 10ms. For constant input current I_{ext}, integrating Eq (66) yields the membrane potential between two spiking events

$$V(t) = V_\infty + (V_0 - V_\infty)e^{-\gamma(t-\hat{t}_0)}, \tag{67}$$

where $\hat{t}_0$ is the time of the most recent spike, $\gamma = g_L/C$ the inverse membrane timescale and $V_\infty = V_R + I_{\mathrm{ext}}/\gamma$ 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

$$p(\hat{t} \in [t, t+\delta t]) = 1 - \exp\left(\int_t^{t+\delta t} \lambda(s)ds\right) \approx \lambda(t)\delta t. \tag{68}$$

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

$$\lambda(t) = \lambda_0 \exp\left(\frac{V(t) - V_T(t)}{\Delta V}\right). \tag{69}$$

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

$$V_T(t) = V_T^* + \sum_{\hat{t}_j < t} \xi(t - \hat{t}_j). \tag{70}$$

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 [43], the following functional form for the adaptation filter was extracted:

$$\xi(s) = \begin{cases} a_\xi & \text{, if } 0 < s \leq T_\xi \\ a_\xi \left(\frac{s}{T_\xi}\right)^{-\beta_\xi} & \text{, if } T_\xi < s < 22\,\text{s}. \end{cases} \tag{71}$$

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.

Model variant with 1s past kernel. For demonstration, we also simulated a variant of the above model with a 1s past kernel

$$\xi^{1\text{s}}(s) = \begin{cases} a_\xi^{1\text{s}} & \text{, if } 0 < s \leq T_\xi \\ a_\xi^{1\text{s}} \left(\frac{s}{T_\xi}\right)^{-\beta_\xi} & \text{, if } T_\xi < s < 1\,\text{s}. \end{cases} \tag{72}$$

All parameters are identical apart from the strength of the kernel $a_\xi^{1\text{s}} = 35.2\,\text{mV}$, which was adapted to maintain a firing rate of 4 Hz despite the shorter kernel.

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 (66) 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 (68). 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.

Table 1. Simulation parameters of the GLIF-ξ model.

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

The parameters were originally extracted from experimental recordings of (n=14) L5 pyramidal neurons [43].

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 (68) 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 $\delta t = 0.5\,\mathrm{ms}$, we first have to transform this to a probability of spiking in the analysis time step $\Delta t = 5\,\mathrm{ms}$. To do so, we compute the probability of no spike in a time step ~~$[t, t+\Delta t]$~~ $[t, t+\Delta t)$ according to

$$p_{\mathrm{sim}}(x_t{=}0|\boldsymbol{x}_t^{-\infty}) = \prod_{j=1}^{\Delta t/\delta t} [1 - \tilde{\lambda}(t + (j-1)\delta t)\delta t], \tag{73}$$

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 (69), but only with respect to past spikes that are emitted at times $\hat{t} < t$. This is because no spike that occurs within ~~$[t, t+\Delta t]$~~ $[t, t+\Delta t)$ 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

$$H_{\mathrm{sim}}(X|\boldsymbol{X}^{-\infty}) = -\frac{1}{N}\sum_{n=1}^{N} \log_2 p_{\mathrm{sim}}(x_{t_n}|\boldsymbol{x}_{t_n}^{-\infty}), \tag{74}$$

and thus the total history dependence

$$R_{\mathrm{tot}} = 1 - \frac{H_{\mathrm{sim}}(X|\boldsymbol{X}^{-\infty})}{\hat{H}(X)}, \tag{75}$$

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~~ thus 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 $R(T, d, \kappa)$. Note that this is a specific property if this model and does not hold in general. For example in experiments, we found that the GLM accounted for less history dependence than model-free estimates (Fig 6).

To map the model neuron to a GLM, we plug the membrane and threshold dynamics of Eq (67) and Eq (70) into the equation for the firing rate Eq(69), i.e.

$$\lambda(t) = \exp\left(\log\lambda_0 + V_\infty - V_T^* + \sum_{\hat{t}_j < t} \xi(t - \hat{t}_j) + (V_0 - V_\infty)e^{-\gamma(t-\hat{t}_0)}\right). \tag{76}$$

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

$$(V_0 - V_\infty)e^{-\gamma(t-\hat{t}_0)} \approx \sum_{\hat{t}_j < t} (V_0 - V_\infty)e^{-\gamma(t-\hat{t}_j)}, \tag{77}$$

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

$$\lambda(t) = \exp\left(\mu \sum_{j=1}^{d} h_j x^-_{t,j}\right), \tag{78}$$

where

$$\mu = \log\lambda_0 + V_\infty - V_T^* \tag{79}$$
$$h_j = \xi(j\delta t) + (V_0 - V_\infty)e^{-\gamma j \delta t}, \tag{80}$$

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

$$\hat{R}_{\mathrm{GLM}}(T, d, \kappa) = 1 - \frac{\hat{H}_{\mathrm{GLM}}(X|\boldsymbol{X}^{-T}_{d,\kappa})}{\hat{H}(X)}. \tag{81}$$

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 81) 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 [46, 47]. Data from mouse primary visual cortex were recorded according to the UK Animals Scientific Procedures Act (1986).

For all recordings, we only analyzed ~~neurons~~ sorted units with firing rates between 0.5 Hz and 10 Hz to exclude the extremes of either inactive ~~neurons or neurons~~ units or units 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 $T_{\text{rec}} \approx 203\,\text{min}$. 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~~ sorted units that satisfied our requirement on the firing rate. More details on the recording procedure can be found in [68], and details on the data set proper can be found in [50].

Salamander retina. Spikes from larval tiger salamander retinal ganglion cells were recorded *in vitro* by extracting the entire retina on an electrode array [69], 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~~ sorted units that satisfied our requirement on the firing rate. More details on the recording procedure and the data set can be found in [48, 49]. The spike recording as obtained from the Dryad database [48].

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~~ sorted units 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 [46, 47]. 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~~ sorted units were located using the location of the electrode contacts provided in [51], and the Allen Mouse Common Coordinate Framework [70]. We analyzed in total $n = 142$ ~~neurons from the rat~~ sorted units from the mouse "Waksman" that belonged to primary visual cortex (irrespective of their layer) and satisfied our requirement on the firing rate. Second, we only selected ~~neurons~~ units 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 [51].

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 $\Delta\kappa_{\min} = 0.01$. In Table 2 we summarize the relevant parameters that were used for embedding optimization.

Details to Fig 3. For Fig 3B, the process was considered for $l = 1$ and an reactivation probability of $m = 0.8$. For $l = 1$, all probabilities can easily be calculated, with marginal probability to be active $p(x_t = 1) = h/(1 - m + mh)$, and conditional

Table 2. Parameters used for embedding optimization.

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)

To facilitate reproduction, we added the settings variable names of the parameters as they are used in the toolbox [37].

probabilities $p(x_t = 1|x_{t-1} = 1) = h + (1 - h)m$ and $p(x_t = 1|x_{t-1} = 0) = h$. From these probabilities, the total mutual information I_{tot} and total history dependence R_{tot} could be directly computed. We then plotted these quantities as a function of h, where values of h were chosen to vary the firing rate between 0.5 and 10 Hz, with a bin size of $\Delta t = 5$ms. For Fig 3C, the binary autoregressive process was simulated for $n = 10^7$ time steps with $m = 0.8$ $(l = 1)$, whereas for $l = 5$, m was adapted to yield approximately the same R_{tot} as for $l = 1$. The input activation probability h was chosen to lead to a fixed probability $p(x = 1) \approx 0.025$, corresponding to 5 Hz firing rate with $\Delta t = 5$ms. Autocorrelation $C(T)$ was computed using the MR.estimator toolbox [53], and $\Delta R(T)$ and $L(T)$ were estimated using plugin estimation. For Fig 3D, the same procedures were applied as in Fig 3C, but now m was varied between 0.5 and 0.95, and h was adapted for each m to hold the firing rate fixed at 5 Hz. For Fig 3E, the same procedures were applied as in Fig 3C, but now l was varied between 1 and 10, and h and m were adapted for each l to hold the firing rate fixed at 5 Hz and R_{tot} fixed at the value for $l = 1$ and $m = 0.8$.

Details to Fig ~~3A~~4A,B. The branching process was simulated using the MR.estimator toolbox, with a time step of $\Delta t = 4$ ms, population rate of 500 Hz and subsampling probability of 0.01. Thus, the subsampled spike train had a firing rate of ≈ 5 Hz. The branching parameter was set to $m = 0.98$ with analytic autocorrelation time $\tau_C(m) = 198$ ms. For a long simulation, autocorrelation $C(T)$ was computed using the MR.estimator toolbox, $L(T)$ using plugin estimation, and $R(T)$ using embedding optimized Shuffling estimator with $d_{\max} = 20$. The generalized timescales τ_R and τ_L were computed with $T_0 = 10$ ms.

Details to Fig 4C,D. The Izhikevich model was simulated with the PyNN toolbox [71], with parameters set to the chattering mode $(a = 0.02, b = 0.2, c = -50, d = 2)$, simulation time bin $dt = 0.01$ ms, and noisy input with mean 0.011 and standard deviation 0.001. For the analysis, a time step of $\Delta t = 1\,\text{ms}$ was chosen. Apart from that, $C(T)$ and $L(T)$ were computed as for Fig 4B. Here, $R(T)$ was computed with BBC and $d_{\max} = 20$, which revealed higher R_{tot} than Shuffling. To compute τ_R, we set $T_0 = 0$.

Details to Fig 4E,F. The GLIF model was simulated as described in Benchmark neuron model (model with 22s past kernel). The analysis time step was $\Delta t = 5\,\text{ms}$. Apart from that, $C(T)$ and $L(T)$ were computed as for Fig 4B. History dependence $R(T)$ was estimated using a GLM as described in Benchmark neuron model. To compute τ_R, we set $T_0 = 10\,\text{ms}$.

Details to Fig 5A,B. In Fig 5A,B, 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 ~~$d \in [1, 100]$~~$d \in [1, 60]$. 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 Fig ~~3C~~5C. In Fig 5C, 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 ~~$d_{\max} = 20$~~up to $d_{\max} = 25$ past bins, 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 Fig ~~4.~~6. For Fig 6, 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 ~~$R(\hat{T}_D)$~~ $R(T_D)$ for the temporal depth ~~$\hat{T}_D$~~ T_D that was estimated with BBC. To optimize the estimate, we computed GLM estimates ~~$\hat{R}(\hat{T}_D)$~~ of $R(T_D)$ 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 ~~$\hat{R}(\hat{T}_D)$~~ $\hat{R}(T_D)$ as a best estimate for R_{tot}. For ~~Figure 3A~~Fig 6A, we plotted only ~~spiketrains~~ spike trains of channels that were identified as single units. For ~~Figure 3B~~Fig 6B, 95% confidence intervals were computed using the standard deviation over $n = 100$ ~~blocks-of-blocks~~ bootstrapping samples. For ~~Figure 3C~~Fig 6C, embedding-optimized estimates with uniform embedding (~~κ~~$\kappa = 0$) were computed with $d_{\max} = 20$ (BBC ~~) or $d_{\max} = 20$~~ and Shuffling) or $d_{\max} = 5$ (Shuffling). Medians were computed over the $n = 28$ ~~neurons in EC~~sorted units in CA1.

Details to Figs ~~5~~ 7 and ~~6.~~8. For Figs 7 and 8, 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 $d_{\max} = 5$. The autocorrelation coefficients $C(T)$ were computed with the MR.Estimator toolbox [53], and the autocorrelation time τ_C was obtained using the `exponential_offset` fitting function. For each recording, we only analyzed the first 40 minutes to have a comparable recording length. For ~~Figure 5~~Fig 7, medians of ~~$\hat{T}_D$ and $\hat{R}_{\text{tot}}$~~ τ_R, τ_C and R_{tot} were computed over all ~~neurons~~ sorted units that were analyzed, and 95% confidence intervals on the medians were obtained by bootstrapping with $n = 10000$ resamples of the median. For ~~Figure 6~~Fig 8, 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 [37], 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~~ 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 7), or even among neurons within a single area (see examples in Fig 8). 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 [20, 21] does not account for higher-order dependencies. We found that the GLM recovered consistently less total history dependence R_{tot} (Fig 6D). 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 $d_{\max} = 5$ past bins with variable exponential scaling κ could account for most of the total history dependence that was estimated with up to $d_{\max} = 20$ bins (Fig 6D). 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 7 and Fig 8) 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 $d_{\max} = 5$ 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 $d_{\max} = 5$ 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 (S1 Fig). 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%) (S1 Fig). The overfitting can, however, be avoided by ~~cross validation~~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 (S1 Fig).

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 total history dependence and ~~temporal depth among neurons~~information timescale across experiments. The recording length affects ~~the estimated~~ estimates of the total history dependence ~~$\hat{R}_{\text{tot}}$~~R_{tot}, and especially ~~the estimated temporal depth $\hat{T}_D$. First, this~~ of the information timescale τ_R. 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~~Moreover, complex embeddings are particular relevant for long past ranges T. Therefore, if ~~the variance is high, smaller temporal depth~~ recordings are shorter, smaller $R(T)$ will be estimated for long past ranges T, leading to smaller estimates of τ_R. We found that for shorter recordings, ~~the estimated total history dependence (thus its amount $\hat{R}_{\text{tot}}$) was~~ estimates of R_{tot} were roughly the same for 90 minutes, but ~~the estimated temporal depth $\hat{T}_D$ was much~~ estimates of τ_R were considerably smaller (S2 and S3 Figs).

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~~ τ_R may not be well resolved. Below 10 minutes, we found that ~~the estimated temporal depth $\hat{T}_D$~~ estimates of τ_R could be less than half of the value that was estimated for 90 minutes, and also ~~the estimated total history dependence $\hat{R}_{\text{tot}}$~~ estimates of R_{tot} showed a notable decrease. In addition, all recordings should have comparable length to prevent that differences in history dependence or ~~temporal depth~~ timescale 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 (). 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.

References

1. Barlow HB. Possible Principles Underlying the Transformations of Sensory Messages. In: Rosenblith WA, editor. Sensory Communication. The MIT Press; 2012. 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. Nature Neuroscience. 2013;16(7):942. doi:10.1038/nn.3431.

4. Atick JJ. Could Information Theory Provide an Ecological Theory of Sensory Processing? Network: Computation in Neural Systems. 1992;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. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. 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. Frontiers in Neuroinformatics. 2014;8. doi:10.3389/fninf.2014.00001.

7. Wibral M, Lizier JT, Priesemann V. Bits from Brains for Biologically Inspired Computing. Frontiers in Robotics and AI. 2015;2. doi:10.3389/frobt.2015.00005.

8. Barlow H. Redundancy Reduction Revisited. Network (Bristol, England). 2001;12(3):241–253.

9. Murray JD, Bernacchia A, Freedman DJ, Romo R, Wallis JD, Cai X, et al. A Hierarchy of Intrinsic Timescales across Primate Cortex. Nature Neuroscience. 2014;17(12):1661–1663. doi:10.1038/nn.3862.

10. Cavanagh SE, Towers JP, Wallis JD, Hunt LT, Kennerley SW. Reconciling Persistent and Dynamic Hypotheses of Working Memory Coding in Prefrontal Cortex. Nature Communications. 2018;9(1):3498. doi:10.1038/s41467-018-05873-3.

11. Wasmuht DF, Spaak E, Buschman TJ, Miller EK, Stokes MG. Intrinsic Neuronal Dynamics Predict Distinct Functional Roles during Working Memory. Nature Communications. 2018;9(1):3499. doi:10.1038/s41467-018-05961-4.

12. Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N. A Hierarchy of Temporal Receptive Windows in Human Cortex. Journal of Neuroscience. 2008;28(10):2539–2550. doi:10.1523/JNEUROSCI.5487-07.2008.

13. 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. Frontiers in Systems Neuroscience. 2018;12. doi:10.3389/fnsys.2018.00055.

14. Wilting J, Priesemann V. Inferring Collective Dynamical States from Widely Unobserved Systems. Nature Communications. 2018;9(1):2325. doi:10.1038/s41467-018-04725-4.

15. Wilting J, Priesemann V. Between Perfectly Critical and Fully Irregular: A Reverberating Model Captures and Predicts Cortical Spike Propagation. Cerebral Cortex. 2019;29(6):2759–2770. doi:10.1093/cercor/bhz049.

16. Zeraati R, Engel TA, Levina A. Estimation of Autocorrelation Timescales with Approximate Bayesian Computations. bioRxiv. 2020; p. 2020.08.11.245944. doi:10.1101/2020.08.11.245944.

17. 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.; 2013. p. 1700–1708. Available from: `http://papers.nips.cc/paper/4873-bayesian-entropy-estimation-for-binary-spike-train-data-using-parame pdf`.

18. Bialek W, Tishby N. Predictive Information. arXiv:cond-mat/9902341. 1999;.

19. Bialek W, Nemenman I, Tishby N. Predictability, Complexity, and Learning. Neural Computation. 2001;13(11):2409–2463. doi:10.1162/089976601753195969.

20. 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. Nature. 2008;454(7207):995–999. doi:10.1038/nature07140.

21. Quinn CJ, Coleman TP, Kiyavash N, Hatsopoulos NG. Estimating the Directed Information to Infer Causal Relationships in Ensemble Neural Spike Train Recordings. Journal of Computational Neuroscience. 2011;30(1):17–44. doi:10.1007/s10827-010-0247-2.

22. Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W. Entropy and Information in Neural Spike Trains. Physical Review Letters. 1998;80(1):197–200. doi:10.1103/PhysRevLett.80.197.

23. Panzeri S, Treves A, Schultz S, Rolls ET. On Decoding the Responses of a Population of Neurons from Short Time Windows. Neural Computation. 1999;11(7):1553–1577. doi:10.1162/089976699300016142.

24. Brenner N, Strong SP, Koberle R, Bialek W, van Steveninck RRdR. Synergy in a Neural Code. Neural Computation. 2000;12(7):1531–1552. doi:10.1162/089976600300015259.

25. Panzeri S, Schultz SR. A Unified Approach to the Study of Temporal, Correlational, and Rate Coding. Neural Computation. 2001;13(6):1311–1349. doi:10.1162/08997660152002870.

26. Stetter O, Battaglia D, Soriano J, Geisel T. Model-Free Reconstruction of Excitatory Neuronal Connectivity from Calcium Imaging Signals. PLoS computational biology. 2012;8(8):e1002653. doi:10.1371/journal.pcbi.1002653.

27. Panzeri S, Treves A. Analytical Estimates of Limited Sampling Biases in Different Information Measures. Network: Computation in Neural Systems. 1996;7(1):87–107. doi:10.1080/0954898X.1996.11978656.

28. Paninski L. Estimation of Entropy and Mutual Information. Neural Computation. 2003;15(6):1191–1253. doi:10.1162/089976603321780272.

29. Panzeri S, Schultz SR, Treves A, Rolls ET. Correlations and the Encoding of Information in the Nervous System. Proceedings of the Royal Society of London Series B: Biological Sciences. 1999;266(1423):1001–1012. doi:10.1098/rspb.1999.0736.

30. Panzeri S, Senatore R, Montemurro MA, Petersen RS. Correcting for the Sampling Bias Problem in Spike Train Information Measures. Journal of Neurophysiology. 2007;98(3):1064–1072. doi:10.1152/jn.00559.2007.

31. Montemurro MA, Senatore R, Panzeri S. Tight Data-Robust Bounds to Mutual Information Combining Shuffling and Model Selection Techniques. Neural Computation. 2007;19(11):2913–2957. doi:10.1162/neco.2007.19.11.2913.

32. Wolpert DH, Wolf DR. Estimating Functions of Probability Distributions from a Finite Set of Samples. Physical Review E. 1995;52(6):6841–6854. doi:10.1103/PhysRevE.52.6841.

33. Nemenman I, Bialek W, de Ruyter van Steveninck R. Entropy and Information in Neural Spike Trains: Progress on the Sampling Problem. Physical Review E. 2004;69(5):056111. doi:10.1103/PhysRevE.69.056111.

34. Archer E, Park I, Pillow J. Bayesian Entropy Estimation for Countable Discrete Distributions. Journal of Machine Learning Research. 2013;15.

35. 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; 2005. p. 1–46. Available from: `https://www.worldscientific.com/doi/abs/10.1142/9789812567772_0001`.

36. Palmigiano A, Geisel T, Wolf F, Battaglia D. Flexible Information Routing by Transient Synchrony. Nature Neuroscience. 2017;20(7):1014. doi:10.1038/nn.4569.

37. Rudelt L, Marx DG, Wibral M, Priesemann V. History Dependence Estimator; 2020. Zenodo. Available from: `https://github.com/Priesemann-Group/hdestimator`.

38. Shannon CE. A Mathematical Theory of Communication. The Bell System Technical Journal. 1948;27(3):379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

39. Goodman J, Weare J. Ensemble Samplers with Affine Invariance. Communications in Applied Mathematics and Computational Science. 2010;5(1):65–80. doi:10.2140/camcos.2010.5.65.

40. Brockwell PJ, Davis RA. Time Series: Theory and Methods. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag; 1991. Available from: `https://www.springer.com/de/book/9780387974293`.

41. Chapeau-Blondeau F. Autocorrelation versus Entropy-Based Autoinformation for Measuring Dependence in Random Signal. Physica A: Statistical Mechanics and its Applications. 2007;380:1–18. doi:10.1016/j.physa.2007.02.077.

42. Albers DJ, Hripcsak G. Using Time-Delayed Mutual Information to Discover and Interpret Temporal Correlation Structure in Complex Populations. Chaos. 2012;22(1):013111–013111–25. doi:10.1063/1.3675621.

43. 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. Journal of Neurophysiology. 2012;107(6):1756–1775. doi:10.1152/jn.00408.2011.

44. Shlens J. Notes on Generalized Linear Models of Neurons. arXiv:14041999 [cs, q-bio]. 2014;.

45. Izhikevich EM. Simple Model of Spiking Neurons. IEEE Transactions on Neural Networks. 2003;14(6):1569–1572. doi:10.1109/TNN.2003.820440.

46. Mizuseki K, Sirota A, Pastalkova E, Buzsáki G. Multi-Unit Recordings from the Rat Hippocampus Made during Open Field Foraging.; 2009. Available from: `http://crcns.org/data-sets/hc/hc-2`.

47. Mizuseki K, Sirota A, Pastalkova E, Buzsáki G. Theta Oscillations Provide Temporal Windows for Local Circuit Computation in the Entorhinal-Hippocampal Loop. Neuron. 2009;64(2):267–280. doi:10.1016/j.neuron.2009.08.037.

48. 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; 2017. Available from: `http://datadryad.org/stash/dataset/doi:10.5061/dryad.1f1rc`.

49. Prentice JS, Marre O, Ioffe ML, Loback AR, Tkačik G, Ii MJB. Error-Robust Modes of the Retinal Population Code. PLOS Computational Biology. 2016;12(11):e1005148. doi:10.1371/journal.pcbi.1005148.

50. Marom S. MEA Data. 2018;1. doi:10.17632/4ztc7yxngf.1.

51. Stringer C, Pachitariu M, Carandini M, Harris K. Eight-Probe Neuropixels Recordings during Spontaneous Behaviors; 2019. Available from: `https://janelia.figshare.com/articles/dataset/Eight-probe_Neuropixels_recordings_during_spontaneous_behaviors/7739750`.

52. Wibral M, Vicente R, Lindner M. Transfer Entropy in Neuroscience. In: Wibral M, Vicente R, Lizier JT, editors. Directed Information Measures in Neuroscience. Understanding Complex Systems. Berlin, Heidelberg: Springer; 2014. p. 3–36. Available from: `https://doi.org/10.1007/978-3-642-54474-3_1`.

53. 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. arXiv:200703367 [physics, q-bio]. 2020;.

54. Wilting J, Priesemann V. 25 Years of Criticality in Neuroscience — Established Results, Open Controversies, Novel Concepts. Current Opinion in Neurobiology. 2019;58:105–111. doi:10.1016/j.conb.2019.08.002.

55. Dong DW, Atick JJ. Temporal Decorrelation: A Theory of Lagged and Nonlagged Responses in the Lateral Geniculate Nucleus. Network: Computation in Neural Systems. 1995;6(2):159–178. doi:10.1088/0954-898X$_{620}$03.

56. Wang XJ, Liu Y, Sanchez-Vives MV, McCormick DA. Adaptation and Temporal Decorrelation by Single Neurons in the Primary Visual Cortex. Journal of Neurophysiology. 2003;89(6):3279–3293. doi:10.1152/jn.00242.2003.

57. Moser EI, Kropff E, Moser MB. Place Cells, Grid Cells, and the Brain's Spatial Representation System. Annual Review of Neuroscience. 2008;31(1):69–89. doi:10.1146/annurev.neuro.31.061307.090723.

58. Masquelier T, Deco G. Network Bursting Dynamics in Excitatory Cortical Neuron Cultures Results from the Combination of Different Adaptive Mechanism. PLOS ONE. 2013;8(10):e75824. doi:10.1371/journal.pone.0075824.

59. Stringer C, Pachitariu M, Steinmetz N, Reddy CB, Carandini M, Harris KD. Spontaneous Behaviors Drive Multidimensional, Brainwide Activity. Science. 2019;364(6437). doi:10.1126/science.aav7893.

60. Meyn SP, Tweedie RL. Markov Chains and Stochastic Stability. Communications and Control Engineering. London: Springer-Verlag; 1993. Available from: //www.springer.com/de/book/9781447132691.

61. Kullback S, Leibler RA. On Information and Sufficiency. The Annals of Mathematical Statistics. 1951;22(1):79–86. doi:10.1214/aoms/1177729694.

62. Kiefer J, Wolfowitz J. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. The Annals of Mathematical Statistics. 1956;27(4):887–906. doi:10.1214/aoms/1177728066.

63. Antos A, Kontoyiannis I. Convergence Properties of Functional Estimates for Discrete Distributions. Random Structures & Algorithms. 2001;19(3-4):163–193. doi:10.1002/rsa.10019.

64. Treves A, Panzeri S. The Upward Bias in Measures of Information Derived from Limited Data Samples. Neural Computation. 1995;7(2):399–407. doi:10.1162/neco.1995.7.2.399.

65. Nemenman I, Shafee F, Bialek W. Entropy and Inference, Revisited. arXiv:physics/0108025. 2001;.

66. Schwarz G. Estimating the Dimension of a Model. Annals of Statistics. 1978;6(2):461–464. doi:10.1214/aos/1176344136.

67. Davison AC, Hinkley DV. Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press; 1997. Available from: https://www.cambridge.org/core/books/bootstrap-methods-and-their-application/ED2FD043579F27952363566DC09CBD6A.

68. Shahaf G, Marom S. Learning in Networks of Cortical Neurons. Journal of Neuroscience. 2001;21(22):8782–8788. doi:10.1523/JNEUROSCI.21-22-08782.2001.

69. Marre O, Amodei D, Deshmukh N, Sadeghi K, Soo F, Holy TE, et al. Mapping a Complete Neural Population in the Retina. Journal of Neuroscience. 2012;32(43):14859–14873. doi:10.1523/JNEUROSCI.0723-12.2012.

70. 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. Cell. 2020;181(4):936–953.e20. doi:10.1016/j.cell.2020.04.007.

71. Davison AP, Brüderle D, Eppler J, Kremkow J, Muller E, Pecevski D, et al. PyNN: A Common Interface for Neuronal Network Simulators. Frontiers in Neuroinformatics. 2009;2. doi:10.3389/neuro.11.011.2008.

Supporting information

S1 Fig. Embedding optimization leads to mild overfitting for short recordings, which can be avoided by cross-validation. Shown is the relative bias ~~, i.e.~~ for two versions of the GLIF model with spike adaption, one with 1s and the other with 22s past kernel. The relative bias refers to 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 $\hat{R}_{\text{tot}}$ was computed by first averaging the relative difference $(\hat{R}(T, d^*, \kappa^*) - R(T, d^*, \kappa^*))/R(T, d^*, \kappa^*)$ for ~~$T \in [\hat{T}_D, T_{\max}]$~~$T \in [T_D, T_{\max}]$, 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 with $d_{\max} = 25$. (Left) For BBC, the relative bias for $\hat{R}_{\text{tot}}$ is zero only if recordings are sufficiently long (≥ 20 minutes for 1s kernel, and ≈ 90 minutes for 22s kernel). 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 one round of cross-validation is applied, i.e. embedding parameters are optimized on ~~one half~~the first, and estimates are computed on the ~~other~~ second half of the data, ~~even for short recordings~~ the bias is approximately zero for BBC even for short recordings, 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 $n = 10$ (45 or 90 min) or $n = 30$ (≤ 20 min) different simulations.

S2 Fig. For the simulated neuron model, recording length has little effect on the estimated total history dependence, but large impact on the estimated ~~temporal depth~~information timescale. (Left) ~~Estimated~~ Mean estimated total history dependence $\hat{R}_{\text{tot}}$ for different recording lengths, relative to the ~~mean~~ true total history dependence ~~estimated for 90 minute recordings (mean over 10 simulations~~R_{tot} of the model (GLIF with spike adaption with 1s or 22s past kernel). As the recording length decreases, ~~also~~ so does $\hat{R}_{\text{tot}}$~~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~~the true R_{tot}. (Right) In contrast, the estimated ~~temporal depth~~ information timescale $\hat{\tau}_R$ decreases strongly with decreasing recording length. With 3 minutes and less, only ~~$\approx 50\%$ of the mean $\hat{T}_D$ for 90 minutes~~ $\approx 75\%$ of the true τ_R is estimated on average. Note that for the simpler 1s model (top), an accurate estimation of the true τ_R is possible for 90 minute recordings, whereas for the 22s model (bottom), the estimated $\hat{\tau}_R$ remains below the true value. Shown are mean values for 30 different simulations for T_{rec} between 1 and 20 minutes, and 10 different simulations for 45 and 90 minutes, as well as 95% confidence intervals on the mean based on bootstrapping.

S3 Fig. ~~Also for~~ For experimental data, too, recording length has little effect on estimated total history dependence, but ~~large~~ larger impact on the estimated ~~temporal depth~~information timescale. (Left) ~~Estimated total history dependence $\hat{R}_{\text{tot}}$~~ Total history dependence R_{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~~ estimate about $\approx 95\%$ as much or more of ~~$\hat{R}_{\text{tot}}$ than~~ R_{tot} as 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~~GLIF model, the estimated ~~temporal depth $\hat{T}_D$ decreases~~ information timescale τ_R decreases more strongly with decreasing recording length. With 10 minutes and more, one estimates around ~~$\approx 50\%$~~ $\approx 75\%$ or more of ~~$\hat{T}_D$~~ the τ_R 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)~~Note that for the experimental data, the estimated timescale of the BBC estimator depends more strongly on the recording time, whereas the Shuffling estimator is more robust, especially for $d_{\max} = 5$. Shown is the median with 95% bootstrapping confidence intervals over $n = 10$ randomly chosen ~~neurons~~ sorted units for each recorded system. Before taking the median over ~~neurons~~sorted units, for each ~~neuron~~ unit 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.

S4 Fig. ~~Estimation~~ Example estimation results for the generalized leaky integrate-and-fire model ~~neuron for simalated example recordings~~(GLIF) with 1s past kernel. For each ~~neuron~~recording length, we show the embedding-optimized estimates of history dependence $R(T)$ ~~for BBC with $d_{\max} = 20$~~ with and without cross-validation, for BBC (red) and Shuffling ~~with $d_{\max} = 20$~~ (blue) with $d_{\max} = 25$, as well as the ground truth for the same embeddings that were found during optimization (dashed lines). ~~Colored dashed~~ Dashed lines indicate the estimated ~~temporal depth $\hat{T}_D$~~ information timescale $\hat{\tau}_R$ and total history dependence $\hat{R}_{\text{tot}}$. Shaded areas indicate ± two standard deviations obtained by bootstrapping.

S5 Fig. ~~Remaining history dependence reveals that many neurons in mouse primary visual cortex still integrate a substantial amount of~~ Example estimation results for the generalized leaky integrate-and-fire model (GLIF) with 22s past ~~information after 80 ms~~kernel. ~~Shown is the remaining history dependence $\Delta\hat{R}(T) \equiv \hat{R}_{\text{tot}} - \hat{R}(T)$ for $T = 80$ ms for all neurons in cortical culture, retina and primary visual cortex~~For each recording length, we show the embedding-optimized estimates of history dependence $R(T)$ with and without cross-validation, for BBC (red) and Shuffling (blue) with $d_{\max} = 25$, 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~~the ground truth for the same embeddings that were found during optimization (dashed lines). Dashed lines indicate the estimated information timescale $\hat{\tau}_R$ and total history dependence $\hat{R}_{\text{tot}}$. Shaded areas indicate ± two standard deviations obtained by bootstrapping.

S6 Fig. Estimation results for all ~~neurons~~ sorted units in rat dorsal hippocampus (layer CA1). For each ~~neuron~~unit, 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 $\hat{T}_D$~~ estimates of the information timescale τ_R and total history dependence ~~$\hat{R}_{\text{tot}}$~~R_{tot}. Also shown is the ~~embedding optimized GLM estimate for the same temporal depth $\hat{T}_D$ as~~ embedding-optimized GLM estimate (violet square) with a past range equal to the temporal depth that was found with ~~BBC~~ the BBC estimator.

S7 Fig. Estimation results for all sorted units in rat cortical culture. For each unit, 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 estimates of the information timescale τ_R and total history dependence R_{tot}. Also shown is the embedding-optimized GLM estimate (violet square) with a past range equal to the temporal depth that was found with the BBC estimator.

~~**S7 Fig. Estimation results for all neurons in rat cortical culture.**~~

S8 Fig. Estimation results for all ~~neurons~~ sorted units in salamander retina. For each unit, 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 estimates of the information timescale τ_R and total history dependence R_{tot}. Also shown is the embedding-optimized GLM estimate (violet square) with a past range equal to the temporal depth that was found with the BBC estimator.

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

S10 Fig. Bootstrapping yields accurate estimates of standard deviation and confidence intervals. (Left) ~~The~~ Shown is the standard deviation on BBC estimates (blue) obtained from 250 “blocks of blocks~~bootstrap samples (Materials and methods)~~” bootstrap samples on a single recording (~~blue)~~GLIF model with 22s past kernel). It 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 ~~$T = \hat{T}_D$~~ $T = T_D$ 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 (CI accuracy). Standard deviation was estimated from 250 “blocks of blocks” bootstrap samples. For each recording length, we computed estimates $\hat{R}$ and the bootstrapping confidence intervals on the 100 simulations~~, and~~. We then computed the confidence level (CI accuracy) by counting how often the true value of R was contained in the estimated confidence interval (green line). Estimates and the true value of R were computed for the same typical embedding parameters d^*, κ^* and $T = T_D$ as before.

S11 Fig. Total history dependence and information timescale for increasing branching parameter m. Similar to the binary autoregressive process, increasing the branching parameter m increases the total history dependence R_{tot}, whereas the information timescale τ_R stays constant, or even decreases for high m. For each m, the input activation probability h was adapted to hold the firing rate fixed at 5 Hz.

S12 Fig. The estimated information timescale varies between estimators. For each sorted unit (grey dots), estimates of the information timescale τ_R are plotted relative to the corresponding BBC estimate for $d_{\max} = 20$. The BBC estimator tends to estimate higher timescales than the Shuffling estimator on recordings of CA1 and cortical culture, whereas for retina the medians of different estimators are more similar. Although estimates of the timescale are highly variable between estimators, Shuffling with only $d_{\max} = 5$ past bins still estimates timescales of at least $80\,\%$ of the timescales that are estimated with BBC. Errorbars indicate median over sorted units and $95\,\%$ bootstrapping confidence intervals on the median.

S13 Fig. Total history dependence and information timescale show no clear dependence on the firing rate, whereas the total mutual information tends to increase with the rate. Shown are the same estimates of the total history dependence R_{tot} and information timescale τ_R as in Fig 7 (Shuffling estimator with $d_{\max} = 5$) versus the firing rates of sorted units (dots). The total mutual information I_{tot} is equal to R_{tot} times the spiking entropy $H(\text{spiking})$ of the respective unit. While I_{tot} tends to increase with firing rate, no clear relation is visible for R_{tot} or τ_R. Errorbars indicate median over sorted units and $95\,\%$ bootstrapping confidence intervals on the median.

S14 Fig. Relationship between total history dependence or information timescale and standard statistical measures of neural spike trains. Estimates of the total history dependence R_{tot} tend to decrease with the median interspike interval (ISI), and to increase with the coefficient of variation C_V. This result is expected for a measure of history dependence, because a shorter median ISI indicates that spikes tend to occur together, and a higher C_V indicates a deviation from independent Poisson spiking. In contrast, the information timescale τ_R tends to increase with the autocorrelation time, as expected, with no clear relation to the median ISI or the coefficient of variation C_V. However, the correlation between the measures depends on the recorded system. For example in retina ($n = 111$), R_{tot} is significantly anti-correlated with the median ISI (Pearson correlation coefficient: $r = -0.69$, $p < 10^{-5}$) and strongly correlated with the coefficient of variation C_V ($r = 0.90$, $p < 10^{-5}$), and τ_R is significantly correlated with the autocorrelation time τ_C ($r = 0.75$, $p < 10^{-5}$). In contrast, for mouse primary visual cortex ($n = 142$), we found no significant correlations between any of these measures. Results are shown for the Shuffling estimator with $d_{\max} = 5$, and $T_0 = 10\,\text{ms}$. Errorbars indicate median over sorted units and $95\,\%$ bootstrapping confidence intervals on the median.

S15 Fig. Excluding short-term contributions helps to differentiate the timescales for different recorded systems. By only considering gains $\Delta R(T)$ for past ranges $T > T_0$ when computing the information timescale τ_R, short-term effects that are related to the refractory period and different firing modes are excluded. The higher T_0, the higher is the distance in the median τ_R between systems (especially between salamander retina and mouse primiary visual cortex). This is because both timescales τ_R and τ_C increase with T_0 for CA1 and primary visual cortex, whereas they decrease for retina. The same holds for the autocorrelation time τ_C, where only delays $T > T_0$ were considered when fitting an exponential decay to the autocorrelograms. Note that if the decay is perfectly exponential, then T_0 does not affect the results. Estimates of R_{tot} and τ_R are shown for the Shuffling estimator with $d_{\max} = 5$. Errorbars indicate median over sorted units and $95\,\%$ bootstrapping confidence intervals on the median.

S16 Fig. Total history dependence decreases for small time bins Δt. The choice of the time bin Δt of the spiking activity has little effect on the information timescale τ_R, whereas the total history dependence R_{tot} decreases for small time bins $\Delta t < 5\,\text{ms}$. This is consistent across experiments. The smaller the time bin, the higher the risk that noise in the spike emission reduces the overall predictability or history dependence in the spiking, whereas an overly large time bin holds the risk of destroying coding relevant time information in the spike train. Thus, we chose the smallest time bin $\Delta t = 5\,\text{ms}$ that does not yet show a substantial decrease in R_{tot}. We do not plot results for higher Δt, because for higher Δt we observed many instances of multiple spikes in the same time bin. Results are shown for the Shuffling estimator with $d_{\text{max}} = 5$, and $T_0 = 10\,\text{ms}$. Errorbars indicate median over sorted units and 95 % bootstrapping confidence intervals on the median.