A Biologically Inspired Attention Model for Neural Signal Analysis

Nicolas Skatchkovsky[1], Natalia Glazman[2], Sadra Sadeh[1, 3, *] and Florencia Iacaruso[1, *]

[1] *The Francis Crick Institute, London, UK.*

[2] *Department of Life Sciences, Imperial College London, London, UK.*

[3] *Department of Brain Sciences, Faculty of Medicine, Imperial College London, London, UK.*

[*] *These authors contributed equally to this work.*

Abstract

Understanding how the brain represents sensory information and triggers behavioural responses is a fundamental goal in neuroscience. Recent advances in neuronal recording techniques aim to progress towards this milestone, yet the resulting high dimensional responses are challenging to interpret and link to relevant variables. In this work, we introduce SPARKS, a model capable of generating low dimensional latent representations of high dimensional neural recordings. SPARKS adapts the *self-attention* mechanism of large language models to extract information from the timing of single spikes and the sequence in which neurons fire using Hebbian learning. Trained with a criterion inspired by predictive coding to enforce temporal coherence, our model produces interpretable embeddings that are robust across animals and sessions. Behavioural recordings can be used to inform the latent representations of the neural data, and we demonstrate state-of-the-art predictive capabilities across diverse electrophysiology and calcium imaging datasets from the motor, visual and entorhinal cortices. We also show how SPARKS can be applied to large neuronal networks by revealing the temporal evolution of visual information encoding across the hierarchy of the visual cortex. Overall, by integrating biological mechanisms into a machine learning model, we provide a powerful tool to study large-scale network dynamics. SPARKS' capacity to generalize across animals and behavioural states suggests it is capable of estimating the internal latent generative model of the world in animals, paving the way towards a foundation model for neuroscience.

Introduction

Recent advancements in both imaging techniques and electrophysiology have enabled neuroscientists to

capture neuronal activity at an unprecedented scale [1]. Experimental scientists now benefit from the capacity to simultaneously record hundreds to thousands, and even a million neurons over extended periods [2], [3], [4], [5]. Interpreting this neural activity and relating it to sensory, motor, and cognitive processes is a key goal of neuroscience [6], yet this task remains inherently challenging for a number of reasons.

First of all, neural recordings are high dimensional by nature, both spatially – because of the number of neurons recorded, and temporally – because of high temporal resolution or long recording times, making them intractable to us. To circumvent this issue, previous studies often relied on single-neuron or average population analysis, and averaged spike trains to consider firing rates [7], [8]. Whilst convenient for analysis, doing so masks the information found at the level of single neurons and single spikes, and relies on arbitrary choices of bin width for averaging. This results in loss of information which can be crucial for analysing neuronal recordings [9], for instance to understand physiological processes with multiple relevant timescales.

Unfortunately, conventional signal processing techniques are ill-suited to operate on sparse, discrete, and non-uniformly sampled spike trains, making them inappropriate for the analysis of neural data. The development of effective mathematical frameworks to analyse such data is a nascent area of research [10]. Additionally, a substantial portion of the variance in neural recordings can typically be attributed to behavioural parameters like uninstructed movements, making it harder to find relevant information pertaining to e.g. sensory processing [11]. Lastly, there is substantial variability across trials, sessions and animals reported by numerous studies when trying to link neural activity to relevant metrics [12], [13]. Altogether, these obstacles render the task of making sense of the high dimensional neuronal responses very challenging.

Focusing on low-dimensional representations of neural data has shown promise, as previous research indicates that information encoded by neural activity in several brain regions and across behavioural tasks can often be represented by a few dimensions [8], [14], [15]. Several machine learning (ML) techniques have been employed to derive meaningful and exploitable low-dimensional representations of high-dimensional neural data [16]. However, these meth

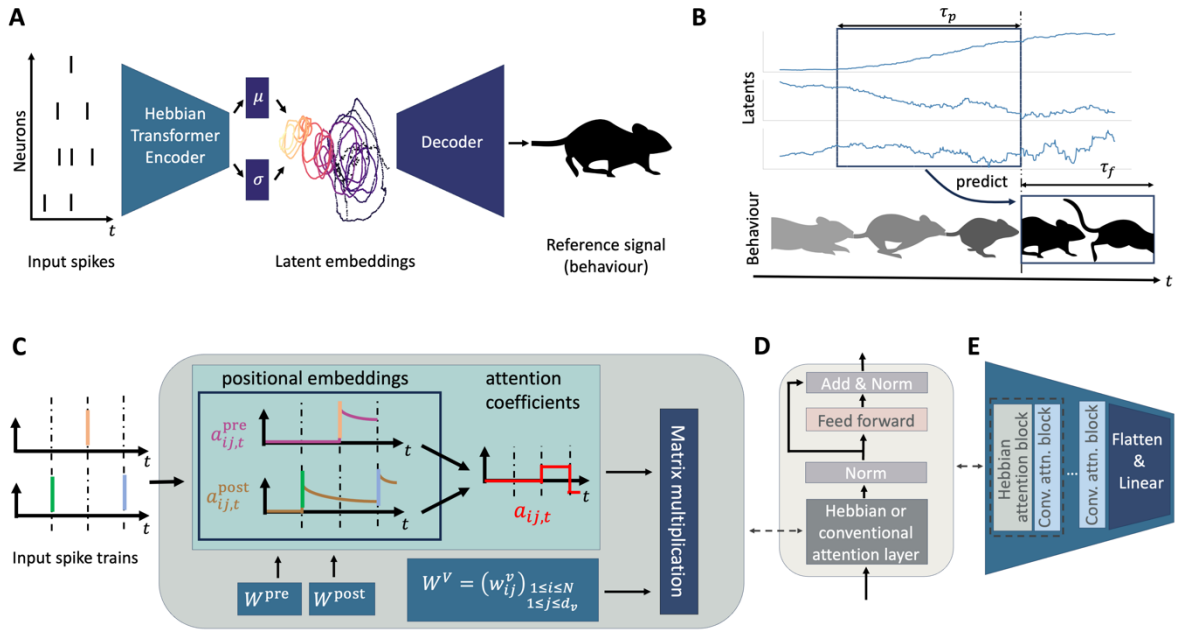

Figure 1 ***Overview of SPARKS.*** *(**A**) Architecture of the proposed variational autoencoder. Blocks* $\boldsymbol{\mu}(\cdot)$ *and* $\boldsymbol{\sigma}(\cdot)$ *represent the mean and standard deviation of the latents distribution. (**B**) Illustration of the predictive conditional distribution: at every time-step, a window* τ_p *of past latent embeddings is used to predict the future* τ_f *samples from the reference signal (**C**) Detailed view of the Hebbian attention layer. (**D**) Schematic of the attention blocks, adapted from [17]. (E) Encoder architecture.*

ods also often ignore precise spike timing and the intrinsic sequential structure of neural data. This leads to loss of critical temporal information and requires extensive pre-processing and manual curation, limiting standardization and reproducibility [7].

To address these challenges, we propose to leverage recent developments in the processing of sequential data via machine learning models, notably with the Transformer model powered by *self-attention* mechanisms [17]. Self-attention consists in computing relationships between parts of an input signal in order for the model to determine which are important for the task. Following their success in natural language processing (NLP) [3] and multimodal data processing [18], Transformers have been recently applied to neural recordings for behaviour prediction, incorporating precise spike times and unit identities [19].

In this work, we propose to use the capabilities of Transformers by designing a self-attention mechanism that leverages single spikes as *tokens*, that is, individual units in the signal. The *Hebbian attention* layer introduced in this work relies on computing relationships between spike trains using the biologically inspired mechanism of Hebbian learning. Our main focus is to obtain expressive and robust latent representations to analyse neural recordings across sessions and animals, with the added benefit of state-of-the-art predictive power. Consequently, we rely on a variational autoencoder (VAE) model [20] to produce the desired

embeddings, rather than a Transformer architecture (Figure 1A). Note that conventional, masked VAEs were recently proposed to model neural distributions [21]. Our approach differs in two ways: firstly, instead of fully connected and conventional layers, the encoder consists of attention blocks, the first of which is the proposed Hebbian attention block. Secondly, we introduce a novel criterion to train the system: taking inspiration from predictive coding, the main objective of the approach is to maximize the probability for the system to predict the correct output for a number of time-steps in the future given the current latent embeddings (Figure 1B, Methods). Importantly, a single model can be used to directly compare the latent embeddings obtained from recordings across sessions or animals within the same latent space.

In this work, we have developed a Sequential Predictive Autoencoder for the Representation of spiKing Signals (SPARKS), and demonstrate that the approach offers several key advantages: training on raw spiking signals or calcium imaging data with minimal preprocessing; directly interpretable low-dimensional latent embeddings; state-of-the-art predictive power from few neurons; unsupervised exploration of the data; and robust encodings across repetitions, sessions, animals, and brain areas. We showcase SPARKS' capabilities using diverse datasets, highlighting its potential to uncover the underlying topology of neuronal representations and its generalisation capabilities across different sessions and animals. We attribute the wide abilities of SPARKS to the use of a Transformer encoder, equipped with our biologically inspired Hebbian attention layer, and training via the proposed objective, which in combination allows us to sequentially extract information at the single spike level with unparalleled precision.

Obtaining Latent Embeddings using a Variational Autoencoder with Hebbian Attention

To obtain consistent latent embeddings that capture most of the variance from high dimensional neuronal responses, SPARKS combines a variational autoencoder with the Hebbian attention layer and predictive learning rule introduced in this work ([10], Figure 1A-B). Following the seminal work [17], the encoder comprises several attention blocks, each composed of an attention layer (Figure 1C), followed by a fully connected feedforward network with residual connections and batch normalisation (Figure 1D-E). The first attention block implements the Hebbian self-attention mechanism that we detail in the next section, with the following blocks implementing conventional dot-product attention ([17], Figure S 1B). To obtain the latent

representation of the input data, the output of the last attention block is flattened and projected into the desired latent dimension d. Throughout, the decoder is a fully connected feedforward neural network, tasked to predict a desired *reference* signal. The reference signal can be chosen as a general supervisory signal such as behavioural data or sensory inputs to perform supervised learning, or as the input spike train itself to perform unsupervised learning. As we will demonstrate, both approaches can be used in complement of each other. When trained via unsupervised learning, our model already captures most of the variance encoded in the neural data, which allows one to freely explore the data using the low-dimensional representations produced by the encoder. This is particularly interesting to investigate how complex combinations of external and internal parameters are encoded in the data. Such an analysis can be completed by training the model via supervised learning to understand the encoding of external parameters that were recorded during the experiments.

We train the proposed autoencoder through a novel objective which is obtained by extending causally conditioned distributions [22], [23]. Causally conditioned distributions model the probability of a random variable (RV) given its own past and that of other RVs of interest: in this context, we maximise the probability to output the desired reference signal given the latent embedding produced by the encoder. More specifically, we introduce a novel criterion based on the *predictive causally conditioned (PCC)* distribution, which consists in maximising the likelihood of future samples from the reference signal given a window of past samples from the latent space (Figure 1B). The reference signal can be chosen either as a relevant timeseries such as behavioural data, or as the input neural signal (Methods). The use of the PCC is inspired by the predictive coding theory, namely that the encodings produced by the generative model in the brain constantly aim at predicting future sensory inputs.

Applying Self-Attention to Neural Data

Self-attention is a key component of the Transformer architecture [17] which consists in computing a similarity measure between different parts of a given input signal to extract contextual information. For instance, in NLP, self-attention measures, for each word in a sentence, the relative importance of the other words to obtain semantic information. How to apply this technique to neural data remains an open question. A growing corpus of works have directly applied traditional attention mechanisms to spiking signals [24],

[25], [26], but they typically relied on the implicit assumption that spike trains were obtained by the encoding of natural, real-valued signals into spikes. Such signals possess specific spatial and/or temporal structures, but it is unclear if such assumptions stand for neural recordings.

To overcome this challenge, we propose a self-attention mechanism for neural data based on a biologically inspired principle: we make use of Hebbian mechanisms, which have long been suggested as one of the key mechanisms underlying learning in biological neural networks [27], [28]. This allows us to apply attention spatially across neurons at every time-instant in the recording, with temporality maintained in the model via the computation of eligibility traces for every pair of neurons, that act as built-in temporal encodings (Figure 1C, Figure S 1A). This is unlike conventional dot-product attention, which relies on various temporal embedding techniques to introduce temporality across tokens presented simultaneously to the model ([17], Figure S 1B).

More specifically, we compute spatial attention between neural sequences as the synaptic coefficients obtained at every time instant via spike-time dependent plasticity (STDP). These coefficients are computed between each pair of neurons from the recording of interest, regardless of prior knowledge of connectivity between neurons. The computation of attention coefficients is carried out by maintaining *eligibility traces* for each pair of neurons, which are calculated using weighted averages of the neurons' spiking history: they are potentiated when either neuron spikes, and decay exponentially otherwise (Figure 1B). The time constant of the exponential decay and the values by which the pre- and post-synaptic traces are updated at spike times could be measured [29] but are *a priori* unknown, and we propose to learn these values via backpropagation individually for each pair of neurons. Our method can also be readily adapted to operate using calcium traces obtained through imaging (Methods). The attention coefficients obtained via STDP are stored in a matrix $\boldsymbol{A}_t \in \mathbb{R}^{N \times N}$, which can be likened to the dot-product between keys and queries in the conventional dot-product attention [17] (Figure 1C, Figure S 1B). These values are then projected into a matrix $\boldsymbol{W}_V$ of dimension $N \times d_v$ where d_v is chosen such that $d_v \ll N$ to reduce computational complexity when the number of neurons is high.

As we will demonstrate throughout, the implementation of an attention mechanism via causal STDP enables the model to capture sequential information. This is key to analysing neural data from brain regions where information is encoded into the precise order of spikes, such as in the entorhinal-hippocampal system during

spatial navigation [30], as highlighted in the following section.

STDP as an Attention Mechanism Captures Sequential Information

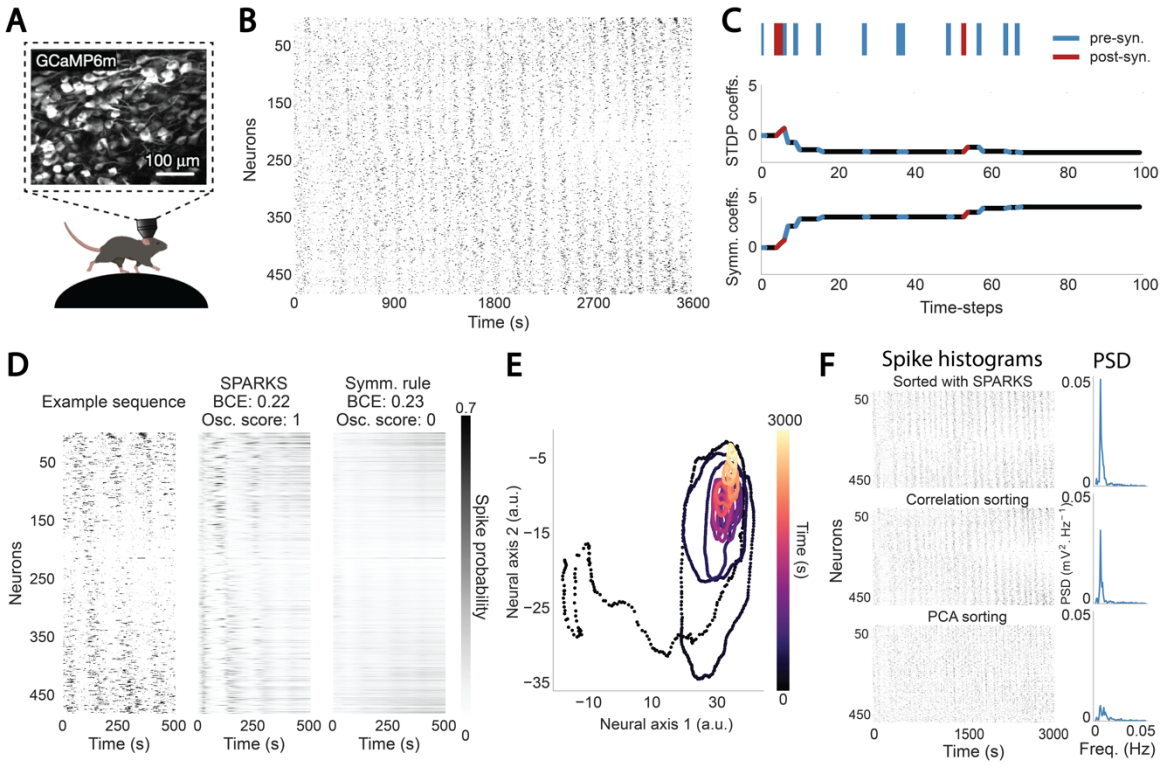

Figure 2 ***SPARKS reveals oscillatory sequences from MEC data.*** *(**A**) Illustration of calcium recording in the MEC of head-fixed mice freely running on a wheel, reproduced from [31]. (**B**) Raster plot of thresholded calcium signals of all cells in an example session (bin size* $= 129$ *ms,* $n = 484$ *cells). Cells were sorted according to their correlation values with one arbitrary cell, in a descending manner. Minutes-long oscillatory sequences appear after sorting. (**C**) Snippet of pre- and post- synaptic firing patterns (top) and corresponding coefficients obtained after application of STDP (middle), and an alternative symmetric rule (bottom). (**D**) Raster plot for left-out data (left) and corresponding predicted spike probabilities after training with SPARKS (centre), and the benchmark symmetric rule (right). (**E**) Two-dimensional embeddings obtained from SPARKS for the whole session after unsupervised training. (**F**) Oscillatory sequences after ordering of the neuron obtained via SPARKS (top), correlation (middle), and PCA sorting (bottom), and corresponding power spectral densities (PSDs, right).*

We begin by motivating the use of STDP over alternative Hebbian rules to capture sequential information in neural data. To do so, we analysed recently published data from the medial entorhinal cortex (MEC), where the authors uncovered oscillatory sequences that are propagated through the MEC in the scale of tens of seconds to minutes (Figure 2A-B, [31]). In such sequences, the information is encoded both spatially and temporally, and analysing them requires keeping track of which neurons fired and in what order. We evaluated the relevance of the proposed attention mechanism by comparing it to an alternative symmetric rule, whereby attention coefficients are potentiated regardless of the order in which pre- and post-synaptic neurons fired (Figure 2C).

We considered a single session recorded with calcium imaging, that lasted roughly 3,600 seconds, and

trained both SPARKS and the alternative rule via unsupervised learning using the first 3,100 seconds of the recording, to test them on the remaining 500 seconds. We found that both rules obtained a similar test binary cross-entropy (BCE) score, but the symmetric rule was unable to recover the oscillatory sequences from the spiking sequence, as measured by the oscillatory score proposed in [31] (Figure 2D, Figure S 2A, Methods). Furthermore, we found that the 2-dimensional embedding of this data produced by SPARKS follow circular trajectories, hence recovering the ring-shaped topology uncovered in [31] (Figure 2E). This was obtained regardless of whether we ordered the neurons prior to training the model, which is necessary to highlight the oscillatory sequences from this data. We were not able to recover such structure from the embeddings produced by the model with the symmetric rule (Figure S 2B). Interestingly, we also found that if the time constant of the exponential decay τ_s and the values by which the pre- and post-synaptic traces are updated at spike times were not learnable, the embeddings produced did not allow us to recover oscillations (Figure S 2C).

From the latent embeddings produced by SPARKS, we were able to retrieve the phase of the oscillations, which allowed us to order neurons according to their correlation to the phase signal (Figure S 2D, Methods), and to obtain clear oscillatory sequences (Figure 2F). To assess the quality of the resulting sorting [31], we calculated the height of the largest peak in the power spectral density (PSD) of the 1-dimensional signal obtained by population decoding of the spike train (Figure 2F, Figure S 2F-G, Methods). SPARKS' sorting had a peak power of 0.05 $\mathrm{mV}^2.\mathrm{Hz}^{-1}$, outperforming the sorting obtained through correlation (0.035 $\mathrm{mV}^2.\mathrm{Hz}^{-1}$) and principal component analysis (PCA, 0.007 $\mathrm{mV}^2.\mathrm{Hz}^{-1}$). Overall, by incorporating STDP to capture temporal information from neuronal recordings, SPARKS can uncover the underlying topology of neuronal representations. In the following sections, we further illustrate the numerous capabilities of the model by using datasets acquired from various brain regions through different recording techniques.

Expressive, Low-Dimensional and Robust Encodings

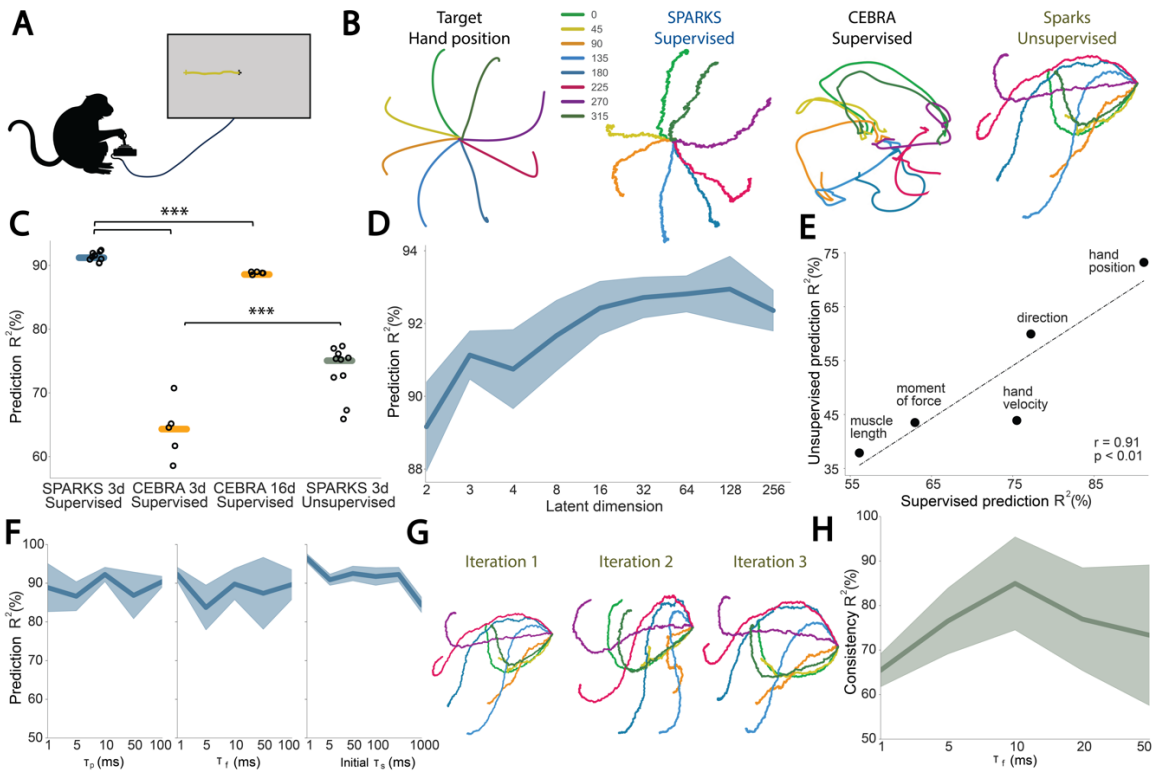

Figure 3 ***Expressive latent embeddings and state-of-the-art prediction from somatosensory cortex data.*** *(**A**) Illustration of a single trial in the reaching task. (**B**) Average hand positions and corresponding latent embeddings obtained through SPARKS supervised, CEBRA supervised and SPARKS unsupervised learning. (**C**) Comparison of R^2 scores for prediction of hand position. Orange lines represent the median over 5 runs, and each dot an individual run. (**D**) Evolution of supervised prediction score with latent space dimensionality. (**E**) Unsupervised vs. supervised prediction scores for the same target variables. Each value represents the mean over* 5 *individual runs. (**F**) Consistency of latent embeddings for supervised learning across hyperparameter values: past window (τ_p), future window (τ_f) and STDP temporal decay (τ_s). (**G**) Example unsupervised latent embeddings across different model initialisations. (**H**) Consistency of latent embeddings for unsupervised learning across prediction window length τ_f. Shaded areas represent standard deviation over* 5 *individual runs. ***: $p < 0.001$, Mann-Whitney U-test. r: Pearson correlation coefficient. p: p-value, t-test.*

Next, we consider data obtained from electrophysiology recordings in Brodmann's area 2 of the somatosensory cortex in monkeys performing a centre-out task (Figure 3A). We use a single recording from the Area2 Bump data set [32], comprising 65 neurons, which was proposed as part of the Neural Latents Benchmark competition [33]. We trained SPARKS via supervised learning to decode the continuous position of the animal's hand, unless mentioned otherwise.

One of our main goals is to produce latent representations of the data that can be directly related to relevant parameters. In this case, the latent embeddings obtained via supervised learning are seen to closely match the trajectories of the animal's hand (Figure 3B). Interestingly, even when training the model via unsupervised learning, we found that the latent trajectories produced by SPARKS are clearly separated across different trial

conditions. The state-of-the-art model [16] was unable to produce embeddings that can be visually matched to the desired reference trajectories, despite providing representations that are clearly separated across trial directions (Figure 3B). To evaluate the role of the proposed attention mechanism, we trained alternative models using the proposed objective. We found that traditional autoencoders did not produce latent embeddings that allowed us to distinguish between different target directions, regardless of the encoder being recurrent or not (Figure S 3A), which we assumed to be necessary to process time-series data. We also trained a model for which we replaced our Hebbian attention layer with conventional dot-product attention. This model produced less visually informative embeddings, notably at the movement onset (Figure S 3A). Alongside the findings from the previous section, these results motivate the use of the Hebbian attention layer introduced in this work to produce low-dimensional representations of neural recordings.

SPARKS also offers very high predictive power, allowing us to obtain highly accurate reconstruction of tasks variables at a single trial level, such as the trajectory of the hand position in this reaching task (Figure 3C). When training the model via supervised learning, predictive power was quantified as the R^2 score obtained using 20% of examples held out for testing (Figure 3C, Methods). This metric is directly obtained from the decoder's outputs, and we obtain $R^2 = 91.2\%$ for the decoding of hand position with $d = 3$ latent dimensions. SPARKS outperformed the state-of-the-art benchmark [16], for 3, and even 16 latent dimensions (Figure 3D, $R^2 = 65\%$, $p < 0.001$; and $R^2 = 88.6\%$, $p < 0.001$, respectively; one-sided Mann-Whitney U-test), as well as all the other models that we implemented (Figure S 3B). For the model proposed in reference [16], predictions were obtained using an auxiliary prediction model, namely a k-nearest neighbour (KNN) regression model. When training SPARKS via unsupervised learning, we can similarly decode hand positions from the latent embeddings via a KNN model. Remarkably, we found that our unsupervised model still achieves higher accuracy than the supervised version of the benchmark, achieving $R^2 = 68.2\%$ with 3 latent dimensions (Figure 3D, $p < 0.001$, one-sided Mann-Whitney U-test). We hence investigated how model parameters affected the prediction capabilities of SPARKS. We found that increasing the length of the past samples window τ_p from 1 to 10 ms improved the performance of the model, but performance plateaued beyond this value (Figure S 3C), hinting at relevant timescales of population code and synaptic transmission. For supervised learning, increasing the value of the prediction window τ_f decreases the prediction performance of the model, although it strongly improved model performance for unsupervised learning (Figure S 3D). We also found that selecting a value of the temporal decay τ_s that is relevant to the timescale

of the behaviour was necessary for the model to perform well, with values in the range of 100 to 500 ms performing best. Finally, increasing the latent dimension slightly increased the accuracy of the model, reaching $R^2 = 93\%$ with a 64-dimensional embedding (Figure 3D), which is relevant to model higher-dimensional behaviours [34]. However, note that the high performance SPARKS consistently achieves with only three latent dimensions is key for obtaining highly informative visualizations of neuronal representations.

In the absence of relevant parameters to perform supervised learning, our model still offers the capacity to uncover the information encoded in the neuronal responses in an unsupervised fashion. To illustrate this, we trained KNN regression models to predict various behavioural variables from the embeddings produced via unsupervised learning. Specifically, we selected the five variables—hand position and velocity, target direction, moment of force applied to the controller, and muscle length—that exhibited the highest unsupervised accuracy. Subsequently, we independently trained our model via supervised learning using each variable as reference signal. We found that the scores obtained from supervised learning exhibit a remarkably high positive correlation with those obtained from unsupervised learning (Pearson correlation coefficient $r = 0.91$, $p < 0.01$, Figure 3E). The unsupervised model hence encodes into the latent space information about the behavioural parameters found in the neural signal, weighted in accordance with their importance in the spike trains. The resulting latent trajectories can be used to investigate biological questions, for instance by exploring which dimensions encode different trial types, outcomes, or sensory information.

Finally, the encodings obtained with SPARKS are robust, both when obtained through supervised and unsupervised learning. To quantitatively evaluate consistency across latent spaces, we trained linear regression models between the embeddings obtained across different repetitions of the training procedure with random initialisations, and evaluated the resulting predictive R^2 score. For supervised learning, the consistency of the embeddings also remains high for all reference signals (Figure S 3E), with a positive correlation between supervised reconstruction accuracy and consistency (Figure S 3F). This suggests that the more informative the neural recording is about the reference signal, the more the latent representation is constrained by the data. We also found very high levels of consistency for a large range of past samples window length τ_p, predictive window τ_f and STDP temporal decay τ_s (Figure 3F), but also other model parameters (Figure S 3G). However, the latent representations derived from distinct behavioural variables

cannot consistently be related to one another via linear regression (Figure S 3E), hinting at the fact that the information encoded for each variable resides within separate lower-dimensional manifolds. When training via unsupervised learning with different random weight initialisations, SPARKS also produces embeddings that are visually similar across repetitions (Figure 3G). Increasing the number of future samples predicted up to $\tau_f = 10$ ms improves the consistency of embeddings across random initialisations (Figure 3H), which can be linked to the optimal value for the parameter τ_p for supervised learning. Taken together, these results demonstrate how SPARKS produces meaningful and robust low-dimensional representations of neural data, providing state-of-the-art performance with only tens of neurons.

Reconstructing Visual Stimuli in Passive Mice

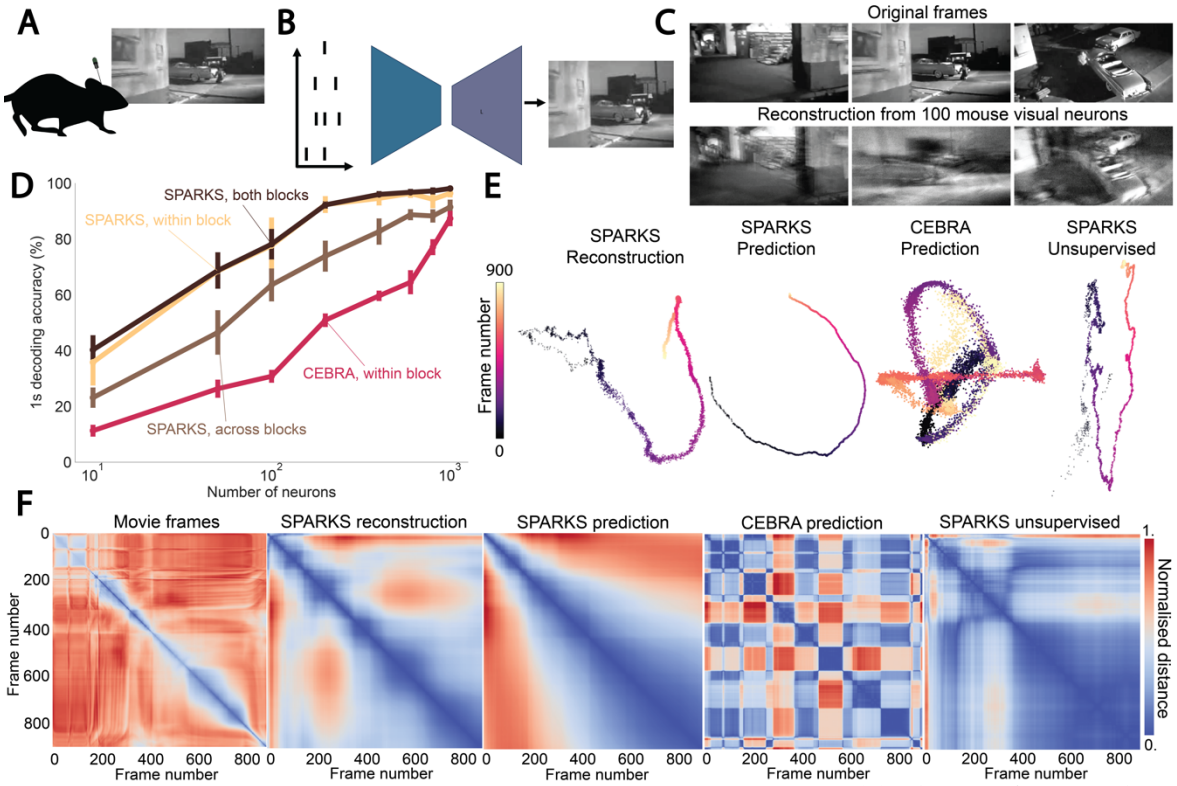

Figure 4 ***Frame reconstruction and prediction from primary visual cortex recordings.*** *(**A**) Schematic of Neuropixels recordings in passive mice viewing a naturalistic film. (**B**) Illustration of the frame reconstruction task: SPARKS was trained to reconstruct actual frames from neural recordings. (**C**) Example frames from the naturalistic movie and corresponding reconstructions using* 100 *neurons recorded from mouse primary visual cortex. (**D**) 1s decoding accuracy as a function of number of neurons for prediction within block (yellow), across two recording blocks separated by 90 minutes (brown), and when training on data from both blocks (black). Within block decoding accuracy is shown for [16] (pink). Error bars represent standard deviations over five repetitions with random sampling of individual units from all recordings. (**E**) Representative examples of two-dimensional embeddings obtained from the neuronal responses to natural movies. From left to right: SPARKS trained for frame reconstruction, SPARKS trained for frame prediction, CEBRA trained for frame prediction, SPARKS trained via unsupervised learning. (**F**) Normalised pairwise distances between frames, computed as inverse, normalised 2-dimensional correlation coefficients; and normalised Euclidian distances between 2-dimensional frame embeddings produced by each model, obtained on a single test example. Same order as in (**D**).*

Next, we asked how our model can be used to decode higher-dimensional stimuli such as natural movie responses, as opposed to low-dimensional parameters such as the direction of the target in the previous task. To this end, we analysed Neuropixels recordings in the primary visual cortex obtained from the Allen visual coding dataset [35]. We focused on passive recordings in mice viewing a 30 second excerpt from the opening of the movie *Touch of Evil* (Figure 4A). In each session, the movie was presented 20 times in two blocks of 10 repetitions, separated by approximately 90 minutes. As illustrated in Figure 4B, we tackled the task of reconstructing individual movie frames from neural activity. To do so, we constructed a pseudo-mouse by stacking the recordings from individual units, chosen at random across all 58 sessions. We started by training SPARKS on the first 9 presentations of the movie from the first block, leaving the last one out for testing, a procedure that we refer to as testing *within block*. We show examples of reconstructed frames in Figure 4C, demonstrating that the model captures high-level scene context, as well as small, fast-moving details (Suppl. Movie 1).

For numerical evaluation, we tested our model on the easier task of predicting frame IDs (i.e., indices ranging from 0-900, 30s movie, 30 fps, Figure 4D). Note that the state-of-the-art reference [16] used as benchmark did not operate directly on the recordings, but first extracted features from the neural data using a Transformer network [16]. Yet, for prediction within block, our model is seen to outperform the best results obtained with the benchmark ([16], 87.4% accuracy, $N = 1{,}000$) using 5 times fewer neurons (SPARKS, 92.3% accuracy, $N = 200$, $p < 0.01$, one-sided Mann-Whitney U-test), suggesting that SPARKS is able to capture key information from the neural data within a small neuronal population.

We also tackled the more challenging task of prediction *across blocks.* To this end, we tested our model on the second set of 10 presentations after training on all examples from the first block, with the two blocks separated by 90 minutes. Analysing neural activity recorded over periods ranging from tens of minutes to hours or more is known to be a difficult task due to a phenomenon known as *neural drift* [36], [37], [38], [39]. Despite these challenges, we found that our model maintains remarkably high accuracy, and even performs as well for predictions across blocks as it does within the same block when $N = 1{,}000$ neurons are used.

In the latent space, SPARKS assigns data points from test examples in the second block close to where

corresponding points from the training set lie in the embedding space (Figure S 4B). The model is hence capable of ignoring the noise caused by neural drift to only extract the meaningful information from the neural data. However, we obtained worse performance and higher variance across repetitions for prediction across blocks when using fewer than 1,000 neurons, indicating that neurons are unevenly impacted by the effects of neural drift across sessions and animals. We also found that training using the first 9 repetitions from both blocks maintains the performance of the model (Figure 4D) compared to training using a single block. This highlights the capacity of SPARKS to find robust latent representations across neural recordings captured over long periods. Additionally, we performed frame prediction using spiking data obtained from pre-processed calcium recordings sessions from the same dataset [40]. We found that SPARKS still performs similarly when decoding information from calcium dynamics (Figure S 4C), establishing our algorithm as the current state-of-the-art for prediction from both types of neural recordings.

To evaluate how the latent representations can give insights about the structure of the input information, in this case the natural scenes, we compared the embeddings obtained through supervised and unsupervised learning after training on the first block using 200 randomly chosen units. Training was carried out via unsupervised learning, or via supervised learning by performing either frame reconstruction or classification. After training via supervised learning, we found that the embeddings produced with both targets form neural trajectories that are closely matched (Figure 4E). Although the neural trajectory obtained via unsupervised learning differs from the other two, we still find that the embeddings of frames from different sections of the movie are clearly distinct in the latent space. To quantitatively evaluate these embeddings, we measured normalised pairwise distances between frame embeddings produced by each model. As a reference, we also computed normalised distances between movie frames (Figure 4E, Methods). We find that in each case, the frame embeddings produced by SPARKS are close in space for frames that are close in time. Remarkably, a structure that can be likened to that of the movie correlations emerges not only for the supervised model, but also for the one trained via unsupervised learning (Figure 4F). Conversely, despite enforcing very strong structure, the embeddings produced by the benchmark model closely position in the latent space frames that are distant in time or correspond to different movie scenes. We were not able to recover such structure from either movie or neural correlations (Figure 4F, Figure S 4D), which implies that previous methods that ignore the sequential nature of neural signals do not allow for their faithful representation.

Learning Across Sessions and Animals

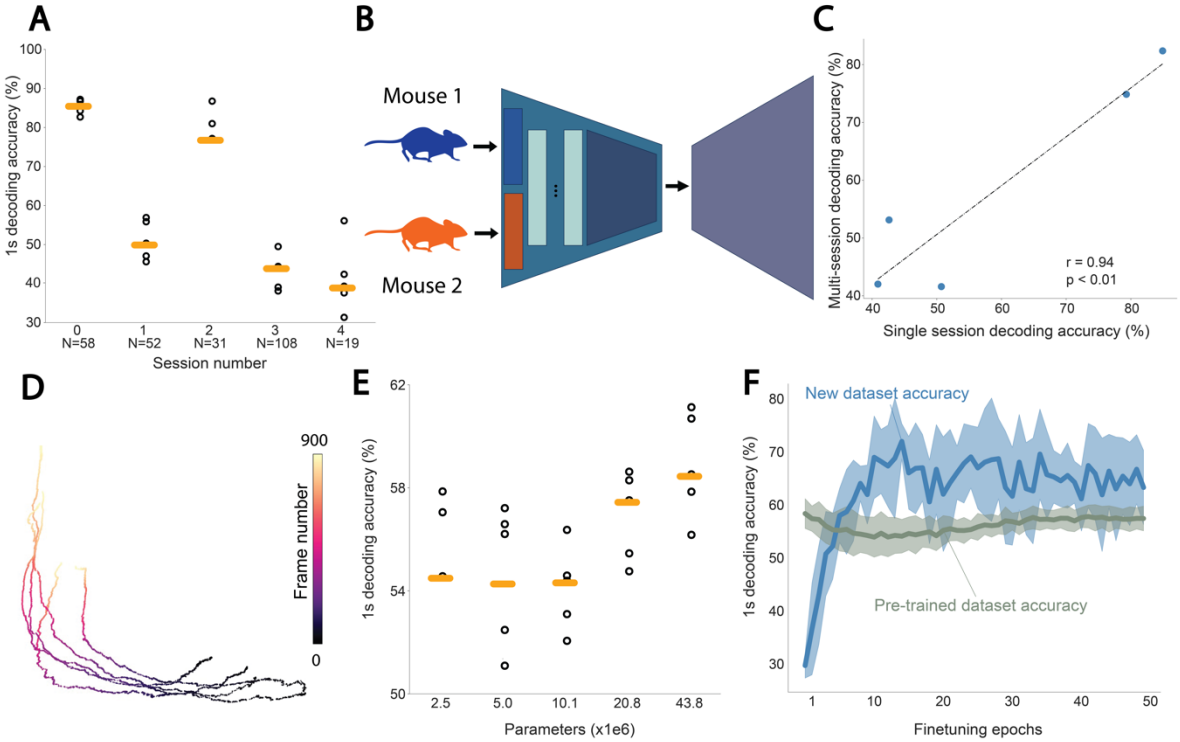

Figure 5 ***SPARKS provides stable embeddings across sessions and animals.*** *(**A**) 1s frame decoding accuracy for prediction of frame IDs with SPARKS for 5 example sessions. Orange lines represent the median over 5 runs, and each dot an individual run. (**B**) Schematic illustrating procedure for training over different recordings with SPARKS. The multisession model is obtained by learning a different Hebbian attention layer for each session while keeping the rest of the network shared across sessions. (**C**) Decoding accuracy of models trained on all sessions and tested on individual sessions versus models trained and tested on single sessions. Each value represents the mean over* 5 *individual runs. (**D**) Two dimensional embeddings corresponding to different sessions obtained from the multi-session model. (**E**) Frame decoding accuracy as a function of model size. Orange lines represent the median over 5 runs, and each dot an individual run. (**F**) Frame decoding accuracy as a function of finetuning epochs. Green trace corresponds to the pre-trained dataset. Blue trace shows accuracy for a newly introduced dataset to an existing model. Shaded areas represent standard deviation over 5 pre-trained models.*

Despite recent developments [5], [41], obtaining stable chronic electrophysiological recordings in mice remains very challenging due to technical limitations such as tissue inflammation and positional changes of the electrodes over time. As a result, researchers often rely on acute recordings. Combining such recordings is common practice and particularly relevant when the brain region of interest allows for the recording of only a few neurons at a time [35], [42]. This approach, however, has the additional confounding effect of masking differences in animal behaviour and state across trials and sessions, even when experimental conditions and stimulation protocols are carefully controlled [43], [44]. Here, we illustrate how SPARKS can tackle this challenge in the task of predicting frames IDs from the Allen Visual dataset. We first trained models using all the neurons from 5 randomly selected sessions, each recorded within a single block. We found a high variance in the predictive accuracy across sessions (Figure 5A). This effect is not explained by the number of neurons in each recording (Figure S 4A, small negative correlation $r = -0.64$, $p > 0.1$),

reinforcing the idea that not all recordings from the same brain region are equally informative. Interestingly, despite the high variance in the predictive power across sessions, the embeddings produced by the model remain highly consistent, as measured by the predictive R^2 obtained after fitting a linear prediction model between the embeddings produced from different recordings ($R^2 > 99\%$ across all pairs, Figure S 4E). To demonstrate robust biological findings, it is necessary to compare embeddings across sessions within the same latent space, which may require extensive work when training the model on single sessions. As discussed, stacking recordings to create pseudo-recordings with hundreds or thousands of units can mitigate this issue. However, it is not always practical to do so, especially for tasks where trials vary in length due to the varying response times of the animals. Combining recordings also ignores varying behavioural and cognitive animal states, such as level of arousal.

To overcome this issue, SPARKS can be employed to encode data from several sessions within the same latent space. This is achieved by learning a different Hebbian attention layer for each session while keeping the rest of the network shared across sessions (Figure 5B). When applying this procedure to the same 5 sessions from the Allen visual dataset, we find that the resulting model maintains predictive accuracy similar to that obtained when training on single sessions, with a very high positive correlation between the two (Figure 5C, $r = 0.94$, $p < 0.01$). Furthermore, we find that the resulting embeddings from each session are closely related (Figure 5D). This demonstrates the capability of our method to find a common latent space which reliably represents the encoding of this natural stimulus within the primary visual cortex. Our results are scalable across model dimensions, which remain small compared to modern LMM architectures (Figure 5E, [45], [46], [47]), hence offering the possibility to train our model without access to the latest GPU architectures.

Finally, analysing new sessions from a previously trained model is straightforward via addition of a new Hebbian attention layer and finetuning (Figure 5F). We find that the model reaches its final accuracy on the new session within $\sim$10 finetuning epochs, with no loss of accuracy when tested on the sessions it was pre-trained on, and reaches similar accuracy to that of a model trained on the single additional session (Figure S 4G). Overall, SPARKS is seen to offer an efficient, powerful and scalable solution to the analysis of neural data across sessions and animals by bringing within the same latent space representations from different datasets. This can then be used to explore how, for example, behavioural variables change the embeddings

across mice.

Uncovering the Functional Hierarchy of Mouse Visual Cortical Areas

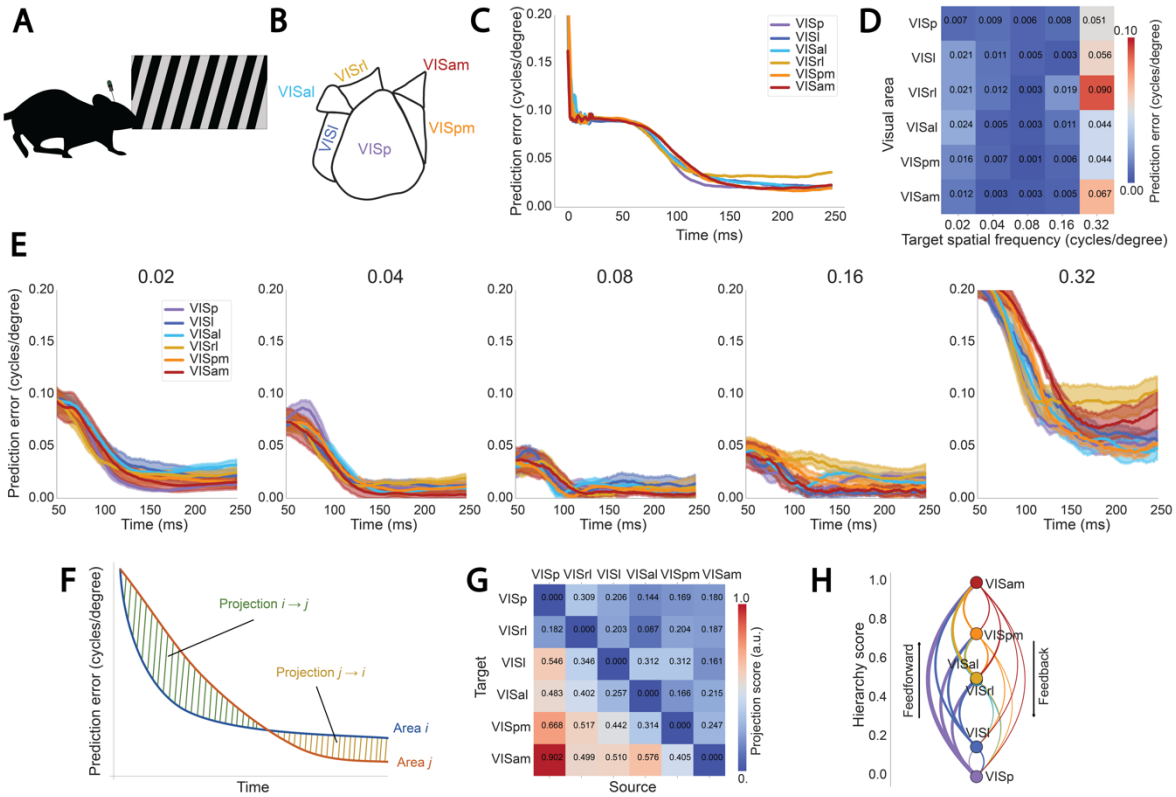

Figure 6 ***SPARKS recovers the hierarchy of the visual cortex.*** *(**A**) Neuropixels recordings in passive mice viewing static gratings. (**B**) Organisation of mice visual cortical areas. (**C**) Evolution of the average spatial frequency prediction error over time for each visual cortical area. Lines represent average over 5 individual runs. (**D**) Average prediction error for each visual cortical area and spatial frequency. (**E**) Evolution of the prediction error over time for each visual cortical area and spatial frequency. Lines represent average over 5 individual runs, shaded areas represent standard deviations. (**F**) Illustration of the computation of projection scores between visual cortical areas. (**G**) Projection scores between all visual cortical areas. (**H**) Functional hierarchy obtained with SPARKS, calculated by computing the difference between their incoming and outgoing projections.*

An important capability of our model is its ability to leverage the temporal dynamics of neuronal data [48] to reveal biological insights that can extend to large neuronal networks. To illustrate this, we studied multi-region interactions from Neuropixels recordings obtained from the mouse primary visual cortex (VISp) and five higher-order visual cortical areas (HVAs), latero-medial area (LM), antero-lateral area (AL), rostro-lateral area (RL), postero-medial area (PM) and antero-medial (AM) from the Allen visual coding dataset (Figure 6A-B) [49]. We trained separate models using 700 neurons recorded in each HVA from awake, head-fixed mice passively presented with static gratings with various spatial frequencies. We trained SPARKS to

predict the spatial frequency of the grating presented to the animals at each time-step with a temporal resolution of 1 millisecond. We focused on spatial frequency because the time course of responses in the primary visual cortex of rodents and primates has been linked to spatial frequency preference, consistent with the coarse-to-fine theory of visual processing [50], [51], [52]. We reasoned that the temporal encoding of spatial frequency across HVAs would be informative of the hierarchical organization of the visual system. As seen in Figure 6C, the models trained using data from all HVAs attain high accuracy across spatial frequencies, with a final average error lower than 0.05 cycles/degree. All of the models start their prediction approximately 50ms from stimulus presentation, but show different dynamics, with VISp reaching its highest accuracy first. We found that models trained on data from visual areas that are higher in the functional hierarchy [49], [53] demonstrate slower dynamics but attain higher accuracies. This general picture hides differences in the areal response to various spatial frequencies: for instance, VISp achieves the lowest error of all areas only for the lowest spatial frequency (Figure 6D). There are also large differences in how quickly gratings are encoded across areas and spatial frequencies; for example, area VISrl provides a very fast but poor estimation of high spatial frequencies, while being more accurate but slower for medium spatial frequencies (Figure 6E).

To uncover dynamic interactions between the different HVAs during bottom-up visual stimulation, we formulated the following hypothesis: an area that generates an accurate representation of a given spatial frequency will selectively transmit this information to the other areas it projects to, provided that these areas have a higher prediction error for the same spatial frequency. In order to remain agnostic and avoid assumptions about connectivity between these areas, we assume that all HVAs are densely connected to each other, and for each pair of areas (i, j), we computed a projection score from area i to area j as the differential integral between the temporal prediction between HVAs i and j (Figure 6E). Considering only positive interactions, we averaged across spatial frequencies and obtained a matrix of projection scores (Figure 6F, Methods). Similarly to previous works [49], [54], we found that VISp is the area with most outgoing projections, while area VISam receives the most inputs from other regions. We then computed hierarchy scores for each area, defined as the difference between incoming projections to the area and outgoing projections from that area. The functional hierarchy we obtained from the temporal encoding of spatial frequencies (Figure 6H) matches the anatomical and functional organization of mouse cortical visual areas previously described [49], [53], demonstrating SPARKS' capability to account for the temporal dynamics of

neuronal signals to reveal large-scale network interactions.

Discussion

Motivated by the advancement of large-scale neuronal recordings and generative machine learning techniques, we introduced SPARKS, a biologically inspired attention model for the analysis of neuronal signals. This new technique can be used to obtain low dimensional embeddings that meaningfully represent the information content from high dimensional data with minimum pre-processing. SPARKS addresses previous limitations by catering to the sequential nature of neuronal signals, and processing information at the single spike level with no need for data processing beyond spike sorting. The high predictive power and latent embeddings obtained with our model allowed us to extract meaningful biological insights from existing data, as validated by revealing temporal dynamics across brain regions, such as the MEC, somatosensory and visual cortex during passive and active behaviours and across different recording techniques.

The core innovation in SPARKS' architecture lies within the Hebbian attention layer, which adapts the highly effective attention mechanism to process spiking signals. Implemented within Transformer architectures, attention is at the centre of recent AI successes, with the development of highly performing large language models (LLMs) and large multimodal models (LMMs) [45], [46], [47], [18]. Attention mechanisms selectively focus on relevant input data, and by assigning weights to relevant inputs, emphasise important features while suppressing irrelevant ones. This enables models to process information more efficiently and effectively, improving performance in various tasks. Conventional attention has been proposed as a model of sparse distributed memory, highlighting the biological relevance of the approach [55]. By taking the inverse approach to derive an attention mechanism from a biological principle, we achieved state-of-the-art performance on various benchmarks, demonstrating the relevance of using the synergies between AI systems development and our biological understanding of intelligence. This is reinforced by the proposed learning objective using predictive causally conditioned distributions, inspired by predictive coding in the brain, which is key to obtaining highly reliable latent embeddings. This is in line with recent work highlighting the benefits for LLMs to simultaneously predict several tokens [56].

Our approach is further justified from the perspective of the Bayesian brain hypothesis, modelling the

autoencoder architecture as a Bayesian encoding-decoding machine [57]. The neural signal measured and forming the inputs $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{r})$ to our model depends on both the animal's internal generative model of the world $p(\boldsymbol{z}|\boldsymbol{r})$, and its sensory inputs $\boldsymbol{r}$. In our model, the encoder performs the approximate posterior inference $p_w(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{r})$ given the neural response to the current sensory inputs and parametric weights $\boldsymbol{w}$. Due to the noise inherent to neural coding and the probabilistic nature of the encoding $\boldsymbol{z}$, the decoding network is a Bayesian decoder tasked to estimate $q_w(\boldsymbol{r}|\boldsymbol{z}, \boldsymbol{x})$. By training encoding and decoding networks together, we obtain latent embeddings that are highly consistent across animals and sessions. This aligns with the Bayesian brain hypothesis that the brain's internal model is calibrated through evolution and exposure to natural stimuli, and evolves slowly during adulthood. We suggest that our model provides an estimate of the animal's latent generative model of the world, providing a new milestone towards a foundation model for neuroscience [58].

Although the memory requirements of the Hebbian attention layer scale quadratically with the number of neurons in the recording, we demonstrated results considering up to $1{,}000$ neurons trained on a single mid-range GPU. By approximating the costly backpropagation-through-time algorithm (Methods), our method is agnostic to the recording duration in terms of memory requirements. As a result, SPARKS will be suitable for analysing electrophysiological data as the scale of these recordings increases with technological advancements. However, the technique cannot yet scale to the latest imaging datasets, comprising up to $1{,}000{,}000$ neurons recorded simultaneously. A limitation of our approach is the computation of attention coefficients between each neuron in the recording, which is not in line with the sparse connectivity measured in the brain. To tackle the challenge of analysing such data, future research directions notably aim to incorporate connectivity data into the model, and to use recent sparse approximations of traditional attention mechanisms [59].

References

[1] N. A. Steinmetz *et al.*, 'Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings', *Science*, vol. 372, no. 6539, p. eabf4588, Apr. 2021, doi: 10.1126/science.abf4588.

[2] C.-H. Yu, J. N. Stirman, Y. Yu, R. Hira, and S. L. Smith, 'Diesel2p mesoscope with dual independent scan engines for flexible capture of dynamics in distributed neural circuitry', *Nat. Commun.*, vol. 12, no. 1, p. 6639, Nov. 2021, doi: 10.1038/s41467-021-26736-4.

[3] J. Demas *et al.*, 'High-speed, cortex-wide volumetric recording of neuroactivity at cellular resolution using light beads microscopy', *Nat. Methods*, vol. 18, no. 9, pp. 1103–1111, Sep. 2021, doi: 10.1038/s41592-021-01239-8.

[4] T. Nöbauer, Y. Zhang, H. Kim, and A. Vaziri, 'Mesoscale volumetric light-field (MesoLF) imaging of neuroactivity across cortical areas at 18 Hz', *Nat. Methods*, vol. 20, no. 4, pp. 600–609, Apr. 2023, doi: 10.1038/s41592-023-01789-z.

[5] A. Yuan, J. Colonell, A. Lebedeva, M. Okun, A. Charles, and T. Harris, ‘Multi-day Neuron Tracking in High Density Electrophysiology Recordings using EMD’, *eLife*, vol. 12, Dec. 2023, doi: 10.7554/eLife.92495.1.

[6] A. E. Urai, B. Doiron, A. M. Leifer, and A. K. Churchland, ‘Large-scale neural recordings call for new insights to link brain and behavior’, *Nat. Neurosci.*, vol. 25, no. 1, pp. 11–19, Jan. 2022, doi: 10.1038/s41593-021-00980-9.

[7] M. Safaie *et al.*, ‘Preserved neural dynamics across animals performing similar behaviour’, *Nature*, vol. 623, no. 7988, pp. 765–771, Nov. 2023, doi: 10.1038/s41586-023-06714-0.

[8] J. A. Gallego, M. G. Perich, S. N. Naufel, C. Ethier, S. A. Solla, and L. E. Miller, ‘Cortical population activity within a preserved neural manifold underlies multiple motor behaviors’, *Nat. Commun.*, vol. 9, no. 1, p. 4233, Oct. 2018, doi: 10.1038/s41467-018-06560-z.

[9] B. Agüera y Arcas, A. L. Fairhall, and W. Bialek, ‘Computation in a single neuron: Hodgkin and Huxley revisited’, *Neural Comput.*, vol. 15, no. 8, pp. 1715–1749, Aug. 2003, doi: 10.1162/08997660360675017.

[10] B. A. Moser and M. Lunglmayr, ‘Spiking Neural Networks in the Alexiewicz Topology: A New Perspective on Analysis and Error Bounds’, May 09, 2023, *arXiv*: arXiv:2305.05772. doi: 10.48550/arXiv.2305.05772.

[11] C. Stringer, M. Pachitariu, N. Steinmetz, C. B. Reddy, M. Carandini, and K. D. Harris, ‘Spontaneous behaviors drive multidimensional, brainwide activity’, *Science*, vol. 364, no. 6437, p. eaav7893, Apr. 2019, doi: 10.1126/science.aav7893.

[12] J. C. Marques, M. Li, D. Schaak, D. N. Robson, and J. M. Li, ‘Internal state dynamics shape brainwide activity and foraging behaviour’, *Nature*, vol. 577, no. 7789, pp. 239–243, Jan. 2020, doi: 10.1038/s41586-019-1858-z.

[13] M. R. Cohen and J. H. R. Maunsell, ‘A Neuronal Population Measure of Attention Predicts Behavioral Performance on Individual Trials’, *J. Neurosci.*, vol. 30, no. 45, pp. 15241–15253, Nov. 2010, doi: 10.1523/JNEUROSCI.2171-10.2010.

[14] M. M. Churchland *et al.*, ‘Neural population dynamics during reaching’, *Nature*, vol. 487, no. 7405, pp. 51–56, Jul. 2012, doi: 10.1038/nature11129.

[15] C. K. Machens, R. Romo, and C. D. Brody, ‘Functional, But Not Anatomical, Separation of “What” and “When” in Prefrontal Cortex’, *J. Neurosci.*, vol. 30, no. 1, pp. 350–360, Jan. 2010, doi: 10.1523/JNEUROSCI.3276-09.2010.

[16] S. Schneider, J. H. Lee, and M. W. Mathis, ‘Learnable latent embeddings for joint behavioral and neural analysis’, Oct. 05, 2022, *arXiv*: arXiv:2204.00673. doi: 10.48550/arXiv.2204.00673.

[17] A. Vaswani *et al.*, ‘Attention is All you Need’.

[18] S. Yin *et al.*, ‘A Survey on Multimodal Large Language Models’, Apr. 01, 2024, *arXiv*: arXiv:2306.13549. Accessed: Jul. 09, 2024. [Online]. Available: http://arxiv.org/abs/2306.13549

[19] M. Azabou *et al.*, ‘A Unified, Scalable Framework for Neural Population Decoding’, Oct. 24, 2023, *arXiv*: arXiv:2310.16046. doi: 10.48550/arXiv.2310.16046.

[20] D. P. Kingma and M. Welling, ‘Auto-Encoding Variational Bayes’, Dec. 10, 2022, *arXiv*: arXiv:1312.6114. doi: 10.48550/arXiv.1312.6114.

[21] A. Schulz *et al.*, ‘Modeling conditional distributions of neural and behavioral data with masked variational autoencoders’, Apr. 25, 2024, *bioRxiv*. doi: 10.1101/2024.04.19.590082.

[22] G. Kramer, ‘Directed information for channels with feedback’, Doctoral Thesis, ETH Zurich, 1998. doi: 10.3929/ethz-a-001988524.

[23] N. Skatchkovsky, O. Simeone, and H. Jang, ‘Learning to Time-Decode in Spiking Neural Networks Through the Information Bottleneck’, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 17049–17059. Accessed: Dec. 04, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/8da57fac3313174128cc5f13328d4573-Abstract.html

[24] M. Yao *et al.*, ‘Attention Spiking Neural Networks’, Sep. 28, 2022, *arXiv*: arXiv:2209.13929. Accessed: Jul. 05, 2023. [Online]. Available: http://arxiv.org/abs/2209.13929

[25] Z. Zhou *et al.*, ‘SPIKFORMER: WHEN SPIKING NEURAL NETWORK MEETS TRANSFORMER’, 2023.

[26] R.-J. Zhu, Q. Zhao, G. Li, and J. K. Eshraghian, ‘SpikeGPT: Generative Pre-trained Language Model with Spiking Neural Networks’, Jun. 26, 2023, *arXiv*: arXiv:2302.13939. Accessed: Jul. 05, 2023. [Online]. Available: http://arxiv.org/abs/2302.13939

[27] D. O. Hebb, *The organization of behavior; a neuropsychological theory*. in The organization of behavior; a neuropsychological theory. Oxford, England: Wiley, 1949, pp. xix, 335.

[28] R. C. Malenka and and R. A. Nicoll, ‘Long-Term Potentiation--A Decade of Progress?’, *Science*, vol. 285, no. 5435, pp. 1870–1874, Sep. 1999, doi: 10.1126/science.285.5435.1870.

[29] G. Q. Bi and M. M. Poo, ‘Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type’, *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 18, no. 24,

pp. 10464–10472, Dec. 1998, doi: 10.1523/JNEUROSCI.18-24-10464.1998.
[30] G. Buzsáki and E. I. Moser, 'Memory, navigation and theta rhythm in the hippocampal-entorhinal system', *Nat. Neurosci.*, vol. 16, no. 2, pp. 130–138, Feb. 2013, doi: 10.1038/nn.3304.
[31] S. Gonzalo Cogno *et al.*, 'Minute-scale oscillatory sequences in medial entorhinal cortex', *Nature*, vol. 625, no. 7994, pp. 338–344, Jan. 2024, doi: 10.1038/s41586-023-06864-1.
[32] R. H. Chowdhury, J. I. Glaser, and L. E. Miller, 'Area 2 of primary somatosensory cortex encodes kinematics of the whole arm', *eLife*, vol. 9, p. e48198, Jan. 2020, doi: 10.7554/eLife.48198.
[33] F. Pei *et al.*, 'Neural Latents Benchmark '21: Evaluating latent variable models of neural population activity', Jan. 17, 2022, *arXiv*: arXiv:2109.04463. doi: 10.48550/arXiv.2109.04463.
[34] J. Manley *et al.*, 'Simultaneous, cortex-wide dynamics of up to 1 million neurons reveal unbounded scaling of dimensionality with neuron number', *Neuron*, vol. 112, no. 10, pp. 1694-1709.e5, May 2024, doi: 10.1016/j.neuron.2024.02.011.
[35] S. E. J. de Vries *et al.*, 'A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex', *Nat. Neurosci.*, vol. 23, no. 1, pp. 138–151, Jan. 2020, doi: 10.1038/s41593-019-0550-9.
[36] H. Lütcke, D. J. Margolis, and F. Helmchen, 'Steady or changing? Long-term monitoring of neuronal population activity', *Trends Neurosci.*, vol. 36, no. 7, pp. 375–384, Jul. 2013, doi: 10.1016/j.tins.2013.03.008.
[37] T. D. Marks and M. J. Goard, 'Stimulus-dependent representational drift in primary visual cortex', *Nat. Commun.*, vol. 12, no. 1, p. 5169, Aug. 2021, doi: 10.1038/s41467-021-25436-3.
[38] L. N. Driscoll, N. L. Pettit, M. Minderer, S. N. Chettih, and C. D. Harvey, 'Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex', *Cell*, vol. 170, no. 5, pp. 986-999.e16, Aug. 2017, doi: 10.1016/j.cell.2017.07.021.
[39] S. Sadeh and C. Clopath, 'Contribution of behavioural variability to representational drift', eLife. Accessed: Jul. 08, 2024. [Online]. Available: https://elifesciences.org/articles/77907
[40] D. Deitch, A. Rubin, and Y. Ziv, 'Representational drift in the mouse visual cortex', *Curr. Biol.*, vol. 31, no. 19, pp. 4327-4339.e6, Oct. 2021, doi: 10.1016/j.cub.2021.07.062.
[41] E. H. van Beest *et al.*, 'Tracking neurons across days with high-density probes', Oct. 16, 2023, *bioRxiv*. doi: 10.1101/2023.10.12.562040.
[42] N. A. Steinmetz, P. Zatka-Haas, M. Carandini, and K. D. Harris, 'Distributed coding of choice, action and engagement across the mouse brain', *Nature*, vol. 576, no. 7786, pp. 266–273, Dec. 2019, doi: 10.1038/s41586-019-1787-x.
[43] A. Lak *et al.*, 'Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon', *eLife*, vol. 9, p. e49834, Apr. 2020, doi: 10.7554/eLife.49834.
[44] S. Musall, M. T. Kaufman, A. L. Juavinett, S. Gluf, and A. K. Churchland, 'Single-trial neural dynamics are dominated by richly varied movements', *Nat. Neurosci.*, vol. 22, no. 10, pp. 1677–1686, Oct. 2019, doi: 10.1038/s41593-019-0502-4.
[45] H. Touvron *et al.*, 'LLaMA: Open and Efficient Foundation Language Models', Feb. 27, 2023, *arXiv*: arXiv:2302.13971. doi: 10.48550/arXiv.2302.13971.
[46] M. Enis and M. Hopkins, 'From LLM to NMT: Advancing Low-Resource Machine Translation with Claude', Apr. 21, 2024, *arXiv*: arXiv:2404.13813. doi: 10.48550/arXiv.2404.13813.
[47] OpenAI *et al.*, 'GPT-4 Technical Report', Mar. 04, 2024, *arXiv*: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.
[48] W. Truccolo, L. R. Hochberg, and J. P. Donoghue, 'Collective dynamics in human and monkey sensorimotor cortex: predicting single neuron spikes', *Nat. Neurosci.*, vol. 13, no. 1, pp. 105–111, Jan. 2010, doi: 10.1038/nn.2455.
[49] J. H. Siegle *et al.*, 'Survey of spiking in the mouse visual system reveals functional hierarchy', *Nature*, vol. 592, no. 7852, pp. 86–92, Apr. 2021, doi: 10.1038/s41586-020-03171-x.
[50] J. Hegdé, 'Time course of visual perception: Coarse-to-fine processing and beyond', *Prog. Neurobiol.*, vol. 84, no. 4, pp. 405–439, Apr. 2008, doi: 10.1016/j.pneurobio.2007.09.001.
[51] P. R. L. Parker *et al.*, 'A dynamic sequence of visual processing initiated by gaze shifts', *Nat. Neurosci.*, vol. 26, no. 12, pp. 2192–2202, Dec. 2023, doi: 10.1038/s41593-023-01481-7.
[52] S. Vreysen, B. Zhang, Y. M. Chino, L. Arckens, and G. Van den Bergh, 'Dynamics of spatial frequency tuning in mouse visual cortex', *J. Neurophysiol.*, vol. 107, no. 11, pp. 2937–2949, Jun. 2012, doi: 10.1152/jn.00022.2012.
[53] J. A. Harris *et al.*, 'Hierarchical organization of cortical and thalamic connectivity', *Nature*, vol. 575, no. 7781, pp. 195–202, Nov. 2019, doi: 10.1038/s41586-019-1716-z.
[54] Q. Wang, O. Sporns, and A. Burkhalter, 'Network analysis of corticocortical connections reveals ventral and dorsal processing streams in mouse visual cortex', *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 32, no. 13, pp. 4386–4399, Mar. 2012, doi: 10.1523/JNEUROSCI.6063-11.2012.
[55] T. Bricken and C. Pehlevan, 'Attention Approximates Sparse Distributed Memory', Jan. 17, 2022,

arXiv: arXiv:2111.05498. doi: 10.48550/arXiv.2111.05498.
[56] F. Gloeckle, B. Y. Idrissi, B. Rozière, D. Lopez-Paz, and G. Synnaeve, 'Better & Faster Large Language Models via Multi-token Prediction', Apr. 30, 2024, *arXiv*: arXiv:2404.19737. doi: 10.48550/arXiv.2404.19737.
[57] R. D. Lange, S. Shivkumar, A. Chattoraj, and R. M. Haefner, 'Bayesian encoding and decoding as distinct perspectives on neural coding', *Nat. Neurosci.*, vol. 26, no. 12, Art. no. 12, Dec. 2023, doi: 10.1038/s41593-023-01458-6.
[58] Y. Zhang *et al.*, 'Towards a "universal translator" for neural dynamics at single-cell, single-spike resolution', Jul. 23, 2024, *arXiv*: arXiv:2407.14668. doi: 10.48550/arXiv.2407.14668.
[59] M. Zaheer *et al.*, 'Big Bird: Transformers for Longer Sequences', Jan. 08, 2021, *arXiv*: arXiv:2007.14062. doi: 10.48550/arXiv.2007.14062.
[60] O. Simeone, 'Machine Learning for Engineers', Higher Education from Cambridge University Press. Accessed: Jul. 11, 2024. [Online]. Available: https://www.cambridge.org/highereducation/books/machine-learning-for-engineers/7FD8622836CAFCF5EDB169E7DC8A1ED4
[61] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, 'Deep Variational Information Bottleneck', Oct. 23, 2019, *arXiv*: arXiv:1612.00410. doi: 10.48550/arXiv.1612.00410.
[62] H. Jang, O. Simeone, B. Gardner, and A. Gruning, 'An Introduction to Probabilistic Spiking Neural Networks: Probabilistic Models, Learning Rules, and Applications', *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 64–77, Nov. 2019, doi: 10.1109/MSP.2019.2935234.
[63] D. P. Kingma and J. Ba, 'Adam: A Method for Stochastic Optimization', Jan. 29, 2017, *arXiv*: arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980.

Methods

Notations

For any two jointly distributed random vectors $\boldsymbol{a} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_T)$ and $b = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_T)$ from time 1 to T, at every time-instant t, we introduce the *predictive causally conditioned (PCC)* distribution of $\boldsymbol{a}$ given $\boldsymbol{b}$ as

$$p^{(\tau_1,\tau_2)}(\boldsymbol{a}_t \,||\, \boldsymbol{b}_t) \;=\; p\big(\boldsymbol{a}_t^{t+\tau_2} \,\big|\boldsymbol{a}^{t-1}, \boldsymbol{b}_{t-\tau_1}^{t}\big). \tag{1}$$

This distribution captures the causal dependence of the τ_2 future samples of RV $\boldsymbol{a}_t^{t+\tau_2} = (\boldsymbol{a}_t, \dots, \boldsymbol{a}_{t+\tau_2})$ on the past τ_1 samples of RV $\boldsymbol{b}_{t-\tau_2}^{t} = (\boldsymbol{b}_{t-\tau_2}, \dots, \boldsymbol{b}_t)$. If $\tau_1 \;=\; T$ and $\tau_2 \;=\; 1$, this recovers the causally conditioned distribution [22].

Problem definition

We study the autoencoder architecture in Figure 1A with the aim of training an encoding network to produce a representation $\boldsymbol{z}$ of the neural input $\boldsymbol{x}$ that is informative about a reference signal $\boldsymbol{r}$. The input $\boldsymbol{x}$ is a neural signal obtained using e.g. electrophysiology or imaging techniques. The reference signal $\boldsymbol{r}$ represents the corresponding target, such as the label corresponding to the behaviour of interest; a set of behavioural variables; or the input itself for unsupervised learning.

To simultaneously optimise encoder and decoder, we consider a data set comprising pairs $(\boldsymbol{x}, \boldsymbol{r})$ of neural inputs $\boldsymbol{x}$ and reference signals $\boldsymbol{r}$, jointly generated from an unknown population distribution $p(\boldsymbol{x}, \boldsymbol{r})$. The inputs are in the form of a discrete collection of vectors over time $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_T)$ with $\boldsymbol{x}_t \in \mathbb{R}^N$, and the reference signals $\boldsymbol{r} = (\boldsymbol{r}_1, \dots, \boldsymbol{r}_T)$, with $\boldsymbol{r}_t \in \mathbb{R}^{N_R}$, define general target signals. Without loss of generality, we assume that all signals are uniformly sampled with a fixed sampling rate dt.

As seen in Figure 1A, the latent representation $\boldsymbol{z}$ is obtained by processing the input signal $\boldsymbol{x}$ through the Hebbian encoder introduced in the main text and for which we provide mathematical details in the following. Given $\boldsymbol{z}$, the system is trained to predict signal $\boldsymbol{r}$ using a decoding network, implemented as an ANN. Following ideas from predictive coding, we maximise the causally conditioned distribution of τ_f future samples from the reference signal $\boldsymbol{r}_t^{t+\tau_f} = (\boldsymbol{r}_t, \dots, \boldsymbol{r}_{t+\tau_f})$ given past input spikes $\boldsymbol{x}^t = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_t)$, averaged over time-steps $t = 1, \dots, T$ as

$$\frac{1}{T}\sum_{t=1}^{T} \log p(\boldsymbol{r}_t^{t+\tau_f} || \boldsymbol{x}_t) = \frac{1}{T}\sum_{t=1}^{T} \log E_{\boldsymbol{z}^t}\, p(\boldsymbol{r}_t^{t+\tau_f} \;|\boldsymbol{z}^t, \boldsymbol{x}^t, \boldsymbol{r}^{t-1}) \,. \tag{2}$$

Note that the evaluation of objective *(2)* is intractable for problems of practical size (see, e.g., [60]) and, as we detail in the following sections, we optimise over the model weights by considering a variational approximation.

Encoding Network

Given the architecture proposed in Figure 1A, the encoder implements a causal stochastic mapping between input $\boldsymbol{x}$ and latent signal $\boldsymbol{z}$ that depends on a vector of parameters $\boldsymbol{w}^e$ as

$$p_{\boldsymbol{w}^e}(\boldsymbol{z}||\boldsymbol{x}) \;=\; \prod_{t=1}^{T} p_{\boldsymbol{w}^e}(\boldsymbol{z}_t \;||\boldsymbol{x}^t),$$

where the distribution $p_{\boldsymbol{w}^e}(\boldsymbol{z}_t||\boldsymbol{x}^t)$ is obtained through the reparametrisation trick as

$$p_{\boldsymbol{w}^e}(\boldsymbol{z}_t||\boldsymbol{x}) \sim \mathcal{N}\left(\boldsymbol{\mu}\big(\boldsymbol{f}_{\boldsymbol{w}^e}(\boldsymbol{x}^t)\big), \boldsymbol{\sigma}^2\big(\boldsymbol{f}_{\boldsymbol{w}^e}(\boldsymbol{x}^t)\big)\right),$$

with $\boldsymbol{f}_{\boldsymbol{w}^e}(\cdot)$ being the output of the encoding network. Samples are obtained as $\boldsymbol{z} = \boldsymbol{\mu}\big(\boldsymbol{f}_{\boldsymbol{w}^e}(\boldsymbol{x})\big) + \boldsymbol{\sigma}^2\big(\boldsymbol{f}_{\boldsymbol{w}^e}(\boldsymbol{x})\big) \odot \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$. The latent signal $\boldsymbol{z}$ is fed to a decoder, whose role is to predict the reference signal $\boldsymbol{r}$. Both blocks $\mu(\cdot)$ and $\sigma(\cdot)$ are implemented as a single linear layer.

Maximum likelihood optimisation

In order to address the maximisation of likelihood *(2)*, we adopt a variational formulation that relies on the introduction of a decoding network that implements at every time-step $t = 1, \dots, T$ the PCC distribution

$$q_{w^d}^{(\tau_p,\tau_f)}\,(\boldsymbol{r}_t \,||\, \boldsymbol{z}_t) \;= q_{\boldsymbol{w}^d}\left(\boldsymbol{r}_t^{\,t+\tau_f} | \boldsymbol{r}^{t-1}, \boldsymbol{z}_{t-\tau_p}^t\right) \tag{3}$$

between latent representation $\boldsymbol{z}$ and reference signal $\boldsymbol{r}$ [23], [61]. The decoder (3) is causal, with memory given by integer $\tau_p \;>\; 1$ and is parametrised by a vector $\boldsymbol{w}^d$.

In practice, we implement the decoder as a model $q_{\boldsymbol{w}^d}\left(\boldsymbol{r}_t^{\,t+\tau_f} \;\middle|\, \boldsymbol{r}^{t-1},\; \boldsymbol{z}_{t-\tau_p}^t\right) := q_{\boldsymbol{w}^d}(\boldsymbol{r}_t^{\,t+\tau_f} \mid \boldsymbol{z}_{t-\tau_p}^t)$, where $q_{\boldsymbol{w}^d}(\boldsymbol{r}_t^{\,t+\tau_f} \mid \boldsymbol{z}_{t-\tau_p}^t)$ is implemented by the output layer of an ANN with inputs given by a window $\boldsymbol{z}_{t-\tau_p}^t$ of samples from the latent representation $\boldsymbol{z}$. Alternatively, one can use an RNN or a probabilistic SNN [62] to directly model the kernel $q_{\boldsymbol{w}^d}\left(\boldsymbol{r}_t^{\,t+\tau_f} \;\middle|\, \boldsymbol{r}^{t-1},\; \boldsymbol{z}_{t-\tau_p}^t\right)$.

With such a network, we use a standard variational inequality [60] to lower bound the decoder log-likelihood at each step $t \;=\; 1, \dots, T$

$$\log p\left(\boldsymbol{r}_{\mathrm{t}}^{\tau_f} \middle| \boldsymbol{x}_{\mathrm{t}}\right) \geq \mathbb{E}_{p_{\boldsymbol{w}^e}\left(\boldsymbol{z}_{t-\tau_p}^t \mid \boldsymbol{x}^t\right)}[\log q_{\boldsymbol{w}^d}^{(\tau_p,\tau_f)}\,(\boldsymbol{r}_t || \boldsymbol{z}_t)] \;+\; \mathrm{KL}(p_{\boldsymbol{w}^e}(\boldsymbol{z}_t \mid \boldsymbol{x}^t) \,||\, p(\boldsymbol{z})).$$

Averaging over time-steps, we obtain the *variational predictive likelihood*

$$\mathcal{L}_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{r}) = \;\mathbb{E}_{p_{\boldsymbol{w}^e}(\boldsymbol{z}||\boldsymbol{x})}\left[\frac{1}{T}\sum_{t=1}^{T} \log q_{\boldsymbol{w}^d}^{(\tau_p,\tau_f)}\,(\boldsymbol{r}_t || \boldsymbol{z}_t)\right] - \frac{1}{T}\mathrm{KL}(p_{\boldsymbol{w}^e}(\boldsymbol{z}||\boldsymbol{r}) \,||\, p(\boldsymbol{z})) \tag{4}$$

where $\boldsymbol{w} \;=\; (\boldsymbol{w}^e, \boldsymbol{w}^{\boldsymbol{d}})$, that we maximise via stochastic gradient descent (SGD) and the Adam optimiser [63] using Monte Carlo gradients.

The gradient with respect to the encoder and decoder weights can be obtained via the stochastic gradient variational Bayes (SGVB) technique [20] as

$$\nabla_{\boldsymbol{w}}\mathcal{L}_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{r}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{p(\boldsymbol{x}^t, \boldsymbol{r}^{t+\tau_f})}\mathbb{E}_{p(\epsilon)}\left[\nabla_{\boldsymbol{w}} \log q_{\boldsymbol{w}^d}\left(\boldsymbol{r}_{\boldsymbol{t}}^{\boldsymbol{t}+\boldsymbol{\tau}_{\boldsymbol{f}}} \middle| \boldsymbol{\mu}\left(\boldsymbol{f}_{\boldsymbol{w}^e}(\boldsymbol{x}^t)\right) + \boldsymbol{\sigma}^2\left(\boldsymbol{f}_{\boldsymbol{w}^e}(\boldsymbol{x}^t)\right) \odot \boldsymbol{\epsilon}\right)\right]. \tag{5}$$

Unbiased estimates of the gradients can be evaluated by using mini-batches $\mathcal{B} \subseteq \mathcal{D}$ from the data set $\mathcal{D}$, along with random samples $\boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, I_{N_Y})$ for $1 \leq \; l \leq \; L$ with $L \geq \; 1$. The gradients $\nabla_{\boldsymbol{w}}$ w.r.t. the

encoder and decoder weights are computed using backpropagation and automatic differentiation tools.

Online optimisation

The computation of gradients (5) is expensive as it requires to apply the backpropagation-through-time algorithm, which scales quadratically in terms of memory consumption with the number of time-steps. To alleviate memory issues during optimisation when the recording duration is large, we propose to use the following approximation to perform SGVB updates at every time-step

$$\nabla_{\boldsymbol{w}}\mathcal{L}_{\boldsymbol{w}}\left(\boldsymbol{x}^t, \boldsymbol{r}_t^{t+\tau_f}\right) \approx \mathbb{E}_{p\left(\boldsymbol{x}^t, \boldsymbol{r}_t^{t+\tau_f}\right)}\mathbb{E}_{p(\epsilon)}\left[\nabla_{\boldsymbol{w}}\log q_{\boldsymbol{w}^d}\left(\boldsymbol{r}_t^{t+\tau_f}\middle|\boldsymbol{\mu}\left(\boldsymbol{f}_{\boldsymbol{w}^e}(\boldsymbol{x}^t)\right) + \boldsymbol{\sigma}^2\left(\boldsymbol{f}_{\boldsymbol{w}^e}(\boldsymbol{x}^t)\right) \odot \boldsymbol{\epsilon}\right)\right],$$

which ignores the contribution of derivatives from previous time-steps $t' < t$.

Hebbian Self-Attention

We propose to apply online STDP at every time-step t to compute a matrix $\boldsymbol{A}_t \in \mathbb{R}^{N_X \times N_X}$ in which each element $a_{ij,t}$ defines an attention coefficient between sequences i and j. We consider an online STDP kernel, whereby coefficients for each synapse $(i \to j)_{1 \le i,j \le N}$ are obtained by maintaining over time a pair of pre- and post-synaptic *eligibility traces* $a_{ij,t}^{\text{pre}}$ and $a_{ij,t}^{\text{post}}$ where $t \in \{1, \dots, T\}$ and N is the number of recorded neurons, Without loss of generality, we assume T to be the temporal horizon of the task. When neurons on either side of the synapse fire, the corresponding traces are updated by an amount determined by learnable weights $\boldsymbol{w}_{ij,t}^{\text{pre}}$ and $\boldsymbol{w}_{ij,t}^{\text{post}}$, and exponentially decay in the absence of a spike by a value determined by scalars τ_{ij}^{pre} and τ_{ij}^{post}. The STDP coefficient $a_{ij,t}$ is potentiated (resp. decreased) by the value of the pre- (resp. post-) synaptic trace when the post- (resp. pre-) synaptic neuron fires. Mathematically, we obtain the coefficient $a_{ij,t}$ as

$$a_{ij,t} := a_{ij,t-1} + s_{i,t} \cdot a_{ij,t}^{\text{post}} - s_{j,t} \cdot a_{ij,t}^{\text{pre}},$$

where $s_{i,t} \in \{0, 1\}$ denotes the state of neuron i at time t, and $a_{ij,t}^{\text{pre}}$ and $a_{ij,t}^{\text{post}}$ are pre- and post-synaptic traces defined as

$$a_{ij,t}^{\text{pre}} := a_{ij,t-1}^{\text{pre}} \cdot \left(1 - \frac{dt}{\tau_{ij}^{\text{pre}}}\right) + w_{ij}^{\text{pre}} \cdot dt \cdot s_{i,t}$$

and

$$a_{ij,t}^{\text{post}} := a_{ij,t-1}^{\text{post}} \cdot \left(1 - \frac{dt}{\tau_{ij}^{\text{post}}}\right) + w_{ij}^{\text{post}} \cdot dt \cdot s_{j,t}.$$

All values $\tau_{ij}^{\text{pre}}, \tau_{ij}^{\text{post}}, w_{ij}^{\text{pre}}$ and w_{ij}^{post} are learnable parameters for each synapse $i \to j$.

Finally, the output of the attention layer is obtained as

$$\text{HebbianAttention}(\boldsymbol{A}_t, \boldsymbol{V}) = \boldsymbol{A}_t \,.\, \boldsymbol{V},$$

where $\boldsymbol{V} \in \mathbb{R}^{N_X \times d_v}$ is a learnable weight matrix defined in analogy to values in conventional dot-product attention [17].

The attention mechanism can further be expanded to account for several versions of the parameter vectors $\mathbf{T}^{\text{pre},(k)} := \left(\tau_{ij}^{\text{pre},(\text{k})}\right)_{ij}$, $\mathbf{T}^{\text{post},(k)}$, $\boldsymbol{W}^{\text{pre},(k)}$, $\boldsymbol{W}^{\text{post},(k)}$, and $\boldsymbol{V}^{(k)}$ for $k = 1, \dots, K$, with K being the number of *attention heads*. The resulting *multi-headed* attention module allows the model to jointly attend to information from different representation subspaces at different positions [17], and can be obtained as

$$\text{MultiHeadedHebbianAttention}_t$$
$$= \boldsymbol{W}_{\boldsymbol{o}}^T \,\text{Concat}(\text{HebbianAttention}(\boldsymbol{A}_t^{(1)}, \boldsymbol{V}^{(1)}), \dots, \text{HebbianAttention}(\boldsymbol{A}_t^{(K)}, \boldsymbol{V}^{(K)}),$$

where $\boldsymbol{W}_{\boldsymbol{o}} \in \mathbb{R}^{Kd_v \times Kd_v}$ is a learnable parameter and $\text{Concat}(\cdot)$ denotes the vertical concatenation of matrices.

Calcium recordings

When considering calcium recording data, we propose to directly use the calcium traces of each neuron as eligibility traces, which removes the heavy processing that comes with spike deconvolution. Mathematically, given inputs $\boldsymbol{x}$, we have

$$a_{ij,t}^{\text{pre}} := \boldsymbol{x}_{i,t},$$

$$a_{ij,t}^{\text{post}} := \boldsymbol{x}_{j,t}$$

and

$$a_{ij,t} := a_{ij,t-1} + a_{ij,t}^{\text{post}} - a_{ij,t}^{\text{pre}}.$$

Mouse MEC dataset

Briefly, activity from MEC cells of head-fixed mice allowed to run on a rotating wheel in darkness was recorded using either two-photon calcium imaging or Neuropixels probes [1]. We consider a single example recording provided by the authors at https://github.com/soledadgcogno/Ultraslow-oscillatory-sequences/tree/main, which included 464 neurons recorded over 3600.51 seconds at a rate of 7.73 Hz. Following the authors' methodology [31], we downsampled the recording by a factor of 4 down to 1.93 Hz. For the prediction task shown in Figure 2, we trained both SPARKS and another autoencoder with the same architecture but implementing the alternative symmetric Hebbian rule on four segments lasting 500 seconds each, the first starting at 600 seconds in the recording, and leaving one out for testing. Results are evaluated numerically through the binary cross-entropy

$$\mathrm{BCE} = \sum_{n,t} r_{n,t} \cdot \log q_{\boldsymbol{w}^d}(r_t \mid \boldsymbol{z}_{t-\tau_p}^t) + (1 - r_{n,t}) \cdot \log(1 - q_{\boldsymbol{w}^d}(r_t \mid \boldsymbol{z}_{t-\tau_p}^t)).$$

We further evaluated the quality of the reconstructions as per the *oscillatory score* proposed in [31]. We show in Figure S 2 the PSD of reconstructions, evaluated using Welch's method with Hamming windows of 66 s (128 bins of 516 ms in each window) and 50% of overlap between consecutive windows. The horizontal lines correspond to the threshold used to evaluate the presence of oscillations [31], which amounts to nine times the average value of the tail of the PSD after the first peak.

Next, we trained SPARKS on the whole recording, starting from 600 seconds in the session as we could not find oscillatory sequences prior to that. From the 2-dimensional latent embeddings hence obtained, we recovered a phase signal by computing at every time-step t the angle between a segment joining the current embedding $\boldsymbol{z_t}$ to the centre of the ring topology, and another joining the embedding $\boldsymbol{z}_t$ to a fixed direction (here, the horizontal, Figure S 2). The phase signal we hence obtain is shown in Figure S 2.

We sorted neurons by ordering them according to their correlation to the phase signal.

Finally, we evaluate the quality of each sorting by population decoding of the sorted signals (Figure S 2) and plotting their PSDs (Figure S 2). We obtain the population decoding $\overline{\boldsymbol{x}}$ of spiking signal $\boldsymbol{x}$ as

$$\overline{x_t} = \frac{1}{N} \sum_i \boldsymbol{x}_{i,t}.$$

Similarly to the oscillatory score proposed in [31], we evaluate the quality of each sorting by measuring the height of the peak above 0 in the PSD.

Macaque reaching task dataset

The dataset corresponds to a single session extracted from [32]. Briefly, electrophysiological recordings were taken from Area 2 of the somatosensory cortex (S1) of a rhesus macaque during an eight-direction center-out reaching task. We consider a session obtained from the Neural Latents Benchmark [33] to train our model. We use 1 ms bin times, and consider recordings from -100 ms to 500 ms after movement onset. Similarly to [33], we consider a 40 ms lag between neural data and movement to align trials. We train SPARKS by considering an encoder with no conventional attention layers, and a 64-dimensional embedding. The decoder is chosen as an MLP with layer dimensions $d \times \tau_p - \frac{d \times \tau_p + 2 \times \tau_f}{2} - N_r \times \tau_f$, where N_r corresponds to the dimension of the reference signal $\boldsymbol{r}$. Models were trained for 500 epochs. Results presented for unsupervised learning were obtained by choosing $\tau_f = 10$.

The alternative model using conventional scaled dot-product attention operates using firing rates smoothed over a 200 ms window, which was chosen after hyperparameter search.

Results from [16] were shared by the authors at https://github.com/AdaptiveMotorControlLab/cebra-figures.

Allen visual coding dataset

The Allen visual coding dataset comprises two-photon calcium imaging and Neuropixels data recorded from the primary visual cortex of mice during presentation of video scenes, and is available to download from the Allen Institute [35]. We consider units sampled at random from all sessions in the Allen visual coding dataset, and stack the raw spike trains to perform movie reconstruction, frame prediction or unsupervised learning.

Each presentation of the excerpt from the movie *Touch of Evil*, performed at a rate of 30 kHz and lasting 30 s, were downsampled with a bin size of 6 ms to speed up training. Calcium recordings were kept to their initial rate of 30 Hz, matching the film frame rate. For frame reconstruction, movie frames $F \in \mathbb{R}^{900 \times h \times w}$ were spatially downsampled to obtain film tensor $\bar{F} \in \mathbb{R}^{900 \times h' \times w'}$ to reduce the computational load of the technique.

The 1-second frame prediction accuracy was computed as

$$\frac{1}{T} \sum_{t=1}^{T} 1(\mathrm{argmax}\big(q(r_t | z^t, x^t)\big) \in [t-6, t+6]),$$

Where $1(\cdot)$ is the indicator function that outputs 1 if its argument is true, and 0 otherwise; and the argmax

function outputs the largest index of its vector argument.

Distances between movie frames were computed as $dist(f_i, f_j) = 1 - corr(f_i, f_j)$ for every pair of frames $(f_i, f_j)_{1 \le i,j \le 900}$ where $corr(\cdot,\cdot)$ is the 2-dimensional correlation normalised to $[0, 1]$.

Distances between the neural activity was computed from the tensor $S \in \mathbb{R}^{10 \times 900 \times 1000 \times 33}$ of spikes over the 10 movie repetitions, for the 900 frames and subsampled at 1kHz, corresponding to 33 time-steps per frame.

The average activity per frame is computed as $\overline{S_f} = \sum_{n,t} S_{nfkt}$. Instantaneous distances between frames are then computed as the $L2$ vector distance $dist\left(\overline{S_{f_i}}, \overline{S_{f_j}}\right) = \sqrt{\sum_k \left(\overline{S_{f_i k}} - \overline{S_{f_j k}}\right)^2}$ and normalised to [0, 1].

Functional Hierarchy of Mouse Visual Cortical Areas

We consider static grating presentations from the Allen visual coding dataset. Gratings were presented for 250 ms in a random order, with spatial frequencies equal to 0.02, 0.04, 0.08, 0.16 and 0.32 cycles per degree. We consider all presentations with a fixed orientation of 150 degrees, and all spatial phases, resulting in 30 training examples and 10 test examples per spatial frequency. Examples were downsampled with a bin size of 1 ms. For each repetition, we used 700 single unit recordings from each visual cortical area (VISp, VISpm, VISam, VISl, VISal and VISrl), which corresponds to the lowest number of neurons collected in a single area across recordings.

We train SPARKS via supervised learning to predict the spatial frequency of the gratings at every time-step, i.e., for all $t = 1, \dots, 250$, we set $r_t \in \{0.02, 0.04, 0.08, 0.16, 0.32\}$. Error at time t for spatial frequency f and cortical area i was computed as the L1 distance

$$e_{i,t}^f = |f - q_{\boldsymbol{w}^d}(r_t \mid \boldsymbol{z}_{t-\tau_p}^t)|.$$

Projections from cortical area i to cortical area j were computed as

$$\text{proj}_{i \to j} = \sum_{f,t} 1\left(e_{i,t}^f < e_{j,t}^f\right) \cdot |e_{i,t}^f - e_{j,t}^f| \,.$$

Code availability

Code: https://www.github.com/FrancisCrickInstitute/sparks (coming soon).

Acknowledgements

Primary funding for this project was provided by the Biotechnology and Biological Sciences Research Council (BBSRC, award ref. BB/X013472/1) and the Francis Crick Institute (award ref. CC2118). The authors wish to thank O. Simeone, A. Schaefer, S. Tootoonian and M. Kollo for their discussions and comments on the manuscript.

Author contributions

F.I. and S.S. co-supervised the work. N.S. developed the algorithm, wrote the code, performed the simulations and wrote the manuscript. N.G. refined the algorithm for calcium data. F.I. and S.S. revised the manuscript.

Materials & Correspondence

Requests should be addressed to N. Skatchkovsky.

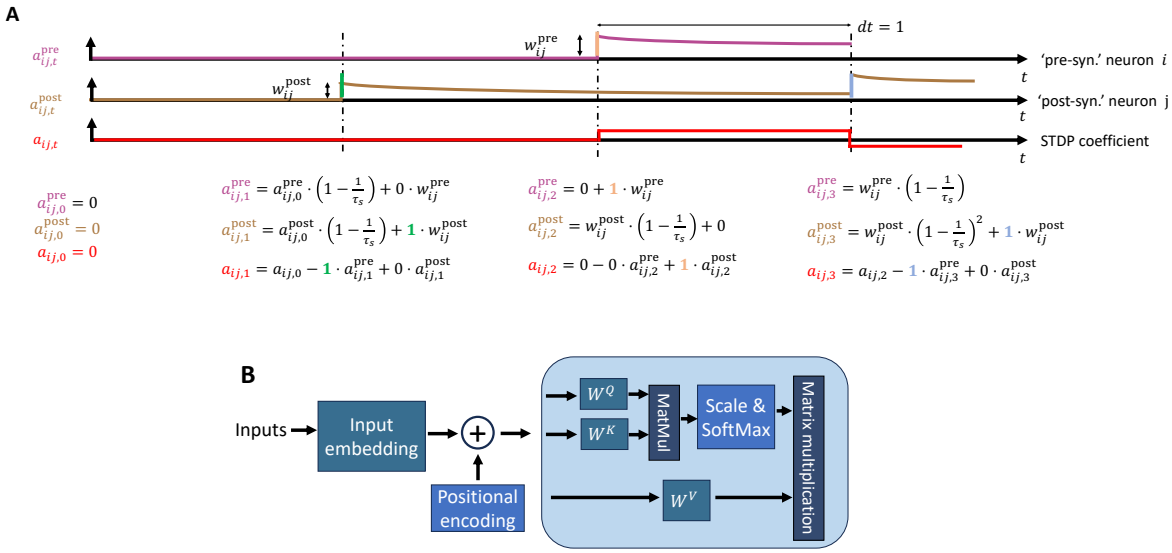

*Figure S 1(**A**) Detailed illustration of the computation of the Hebbian attention coefficients via STDP. (**B**) Schematic of the conventional dot-product attention layer, adapted from [17].*

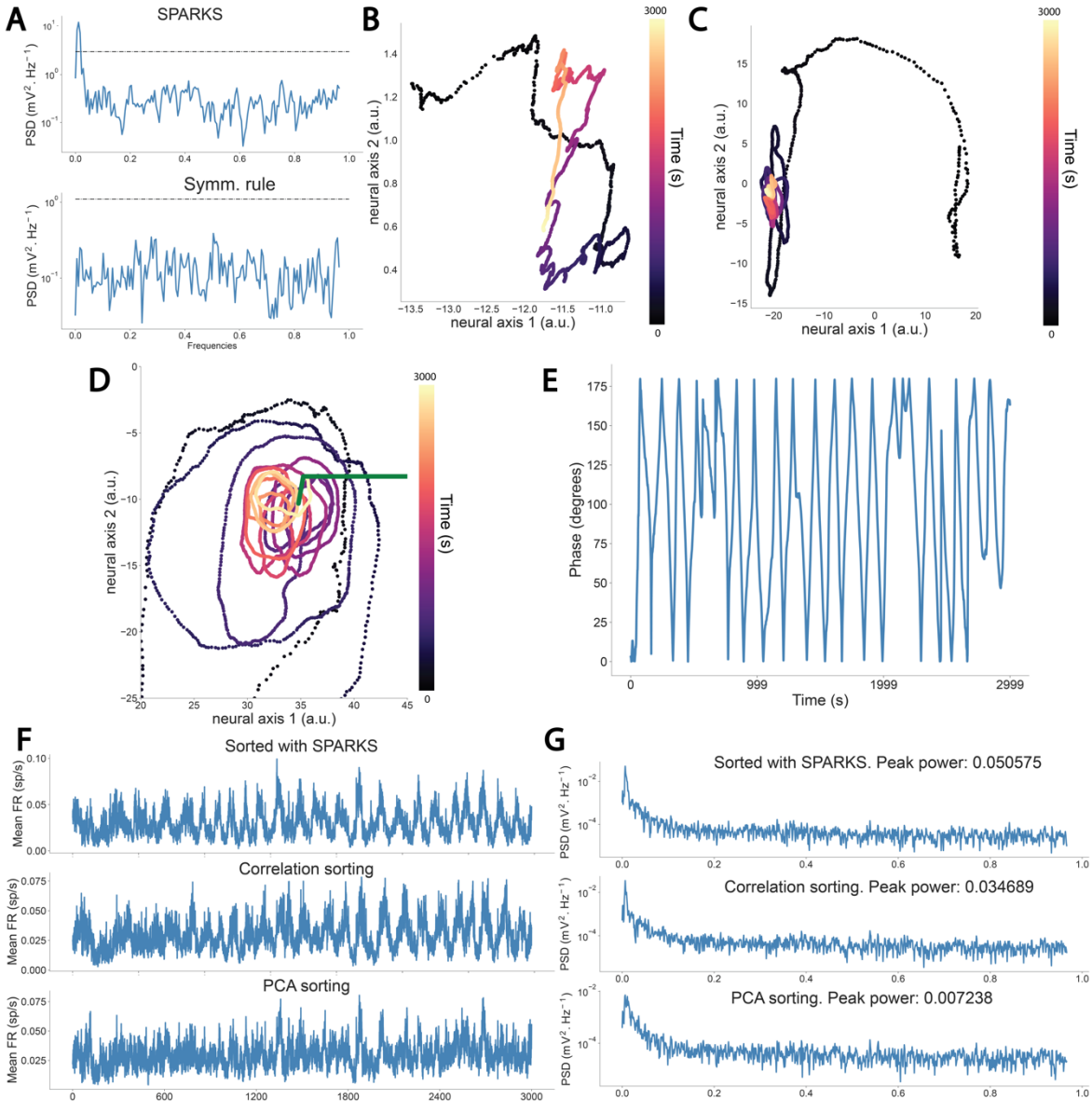

Figure S 2 ***(A)*** *Power spectral density (PSD) of the phase of the oscillation of the signal, as defined in [31].* ***(B)*** *Latent embeddings obtained with the alternative symmetric rule.* ***(C)*** *Latent embeddings obtained with non-learnable parameters* $\tau_{ij}^{pre}, \tau_{ij}^{post}, w_{ij}^{pre}$ *and* w_{ij}^{post} *for all synapses* $i \to j$*.* ***(D)*** *Illustration of the angle used to recover the phase of the oscillations from the latent embeddings.* ***(E)*** *Phase signal recovered from the latent embeddings.* ***(F)*** *Oscillatory signals recovered by population decoding of the spike trains with different sortings of the neurons.* ***(G)*** *Corresponding PSDs.*

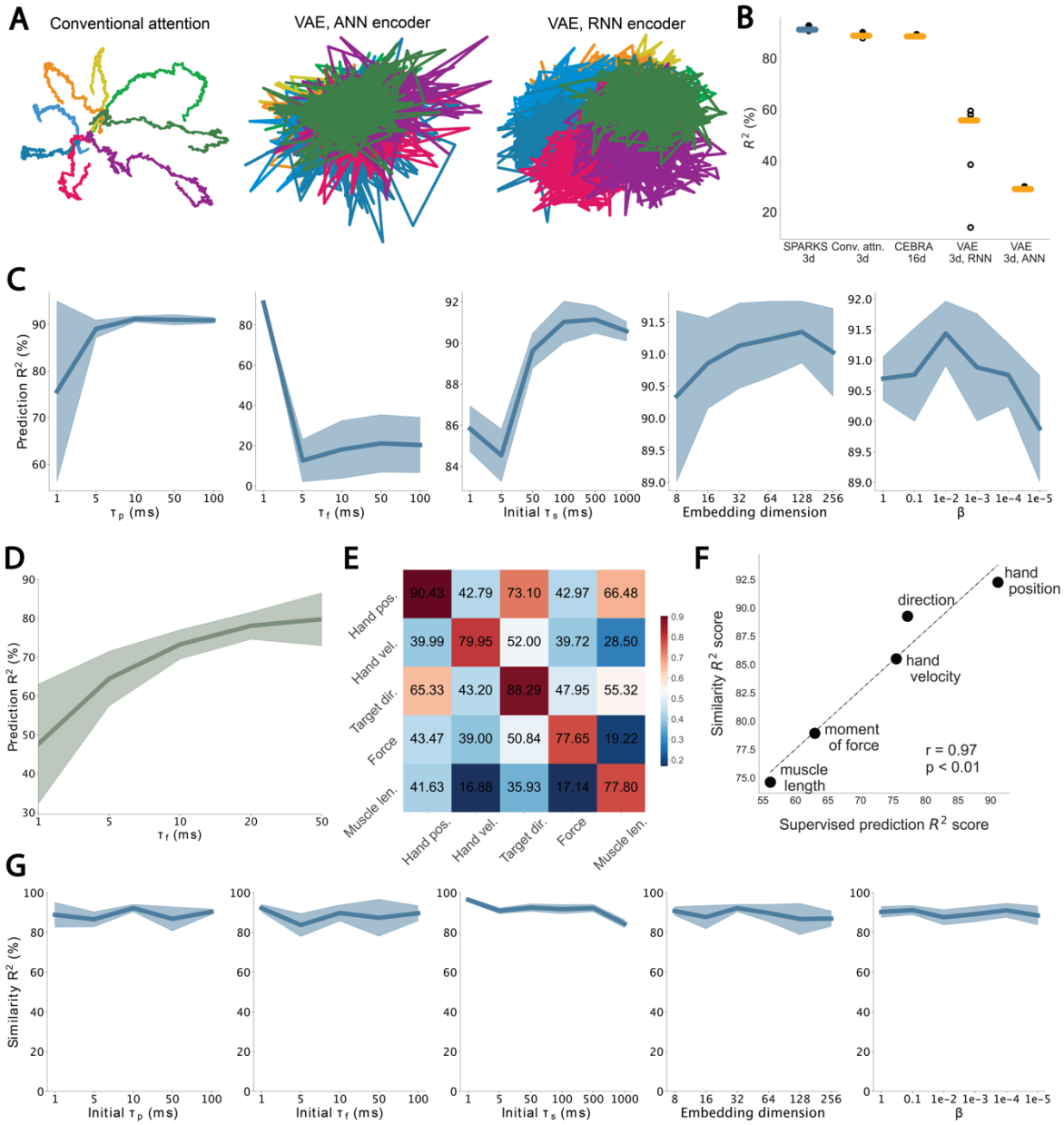

Figure S 3 ***(A)*** *Example latent embeddings obtained from training models to decode the position of the animal's hand. From left to right: Alternative model with conventional dot-product attention, VAE with a traditional ANN encoder, VAE with a recurrent neural network (RNN) encoder.* ***(B)*** *Comparison of R^2 scores for prediction of hand position. Orange lines represent the median over 5 runs, and each dot an individual run. From left to right: SPARKS with a 3-dimensional latent space, alternative model with conventional dot-product attention for a 3-dimensional latent space, CEBRA with a 16-dimensional latent space, VAE with a traditional ANN encoder for a 3-dimensional latent space, VAE with a recurrent neural network (RNN) encoder for a 3-dimensional latent space.* ***(D)*** *Prediction R^2 scores for unsupervised learning across prediction window length τ_f. Shaded areas represent standard deviation over 5 individual runs.* ***(E)*** *Similarity R^2 scores for supervised learning across different reference signals.* ***(F)*** *Consistency vs. prediction R^2 scores for the same target variables. Each value represents the mean over 5 individual runs.* ***(G)*** *Consistency of latent embeddings for supervised learning across hyperparameter values. r: Pearson correlation coefficient. p: p-value, t-test.*

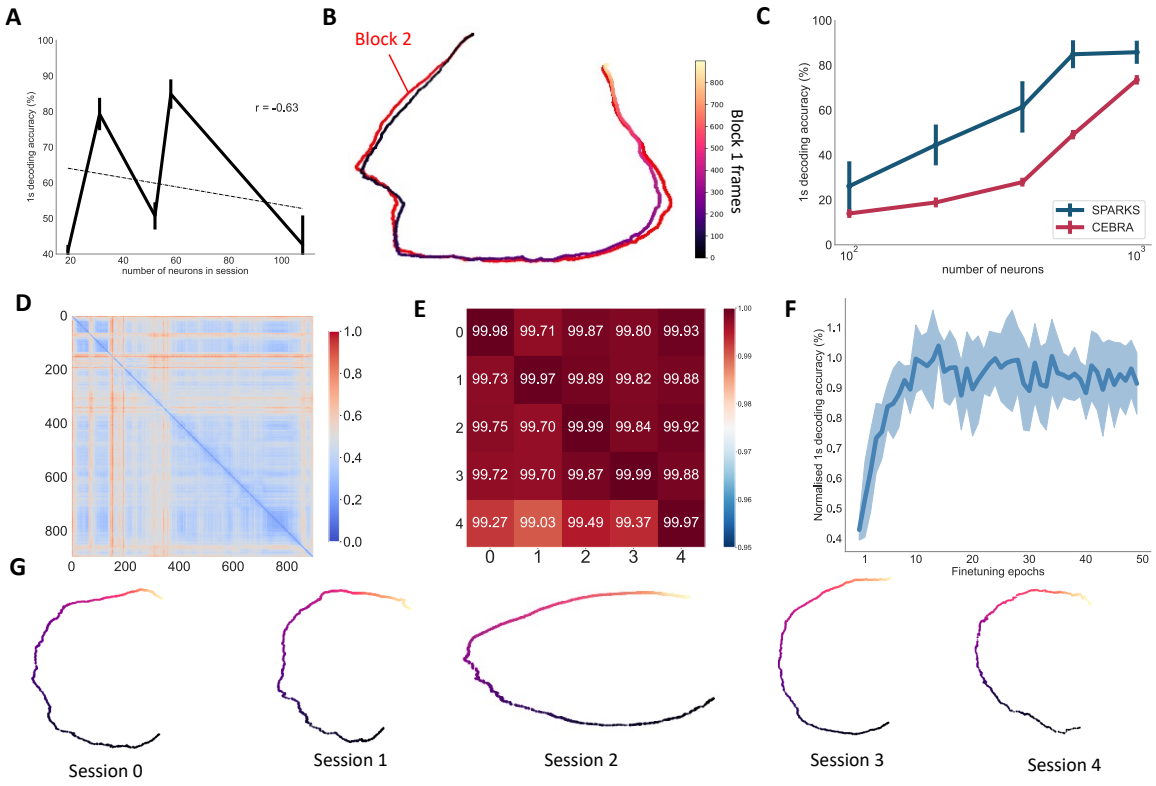

Figure S 4 ***(A)*** *1s decoding accuracy as a function of number of neurons in a session. Error bars represent standard deviation across 5 repetitions. r: Pearson correlation coefficient.* ***(B)*** *Latent embeddings averaged across examples from blocks 1 and 2.* ***(C)*** *1s decoding accuracy as a function of number of randomly sampled neurons. Error bars represent standard deviation across 5 repetitions.* ***(D)*** *Normalised Euclidian distance between the average firing rate per frame in block 1 across 1,000 neurons.* ***(E)*** *Similarity R^2 scores between the latent embeddings obtained for different example sessions, averaged over 5 repetitions and 10 movie presentations from block 1.* ***(F)*** *Evolution of the 1s decoding accuracy during fine-tuning on a single session, normalised to the decoding accuracy of a model trained on this session only.* ***(G)*** *Latent embeddings averaged across examples from blocks 1 for 5 example sessions.*