A Biologically Inspired Attention Model for Neural Signal Analysis

Nicolas Skatchkovsky[1], Natalia Glazman[2], Alexander Egea-Weiss[3], Sadra Sadeh[1, 4, *] and Florencia Iacaruso[1, *]

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

[2] *Department of Biomedical Computing, King's College London, London, UK.*

[3] *Department of Biomedicine, University of Basel, Basel, Switzerland.*

[4] *Centre for Developmental Neurobiology, King's 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. Although existing machine learning models propose to do so, they often sacrifice interpretability for predictive power, effectively operating as black boxes. In this work, we introduce SPARKS, a biologically inspired model capable of high decoding accuracy and interpretable discovery within a single framework. 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 low-dimensional latent embeddings that are robust across sessions and animals. By directly capturing the underlying data distribution through a generative encoding-decoding framework, SPARKS exhibits state-of-the-art predictive capabilities across diverse electrophysiology and calcium imaging datasets from the motor, visual and entorhinal cortices. Crucially, the Hebbian coefficients learned by the model are interpretable, allowing us to infer the effective connectivity and recover the known functional hierarchy of the mouse visual cortex. Overall, SPARKS unifies representation learning, high-performance decoding and model interpretability in a single framework by bridging neuroscience and AI, providing a powerful and versatile tool for dissecting neural computations and marking a step towards the next generation of biologically inspired intelligent systems.

Introduction

Recent advancements in imaging and electrophysiology techniques have enabled neuroscientists to capture neuronal activity at an unprecedented scale [1], yielding recordings comprising 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 challenging for a number of reasons. Neural recordings are high dimensional, both spatially – because of the number of neurons recorded, and temporally – because of high temporal resolution or long recording times, making them intractable. Beyond technical limitations, linking the information in neuronal recordings to relevant sensory or cognitive processes such as decision making is inherently difficult because a substantial portion of the variance can typically be attributed to behavioural parameters like uninstructed movements [7]. As a result, there is a substantial amount of variability across trials, sessions and animals reported by numerous studies when trying to link neural activity to relevant metrics [8], [9]. These differences can also result from separate neuronal populations being recorded in different sessions, from changes in neural activity due to different brain states, or from phenomena such as neural drift [10]. Developing tools to tackle these challenges is now imperative to the field.

Finding suitable low-dimensional representations is a promising avenue to understand complex, high-

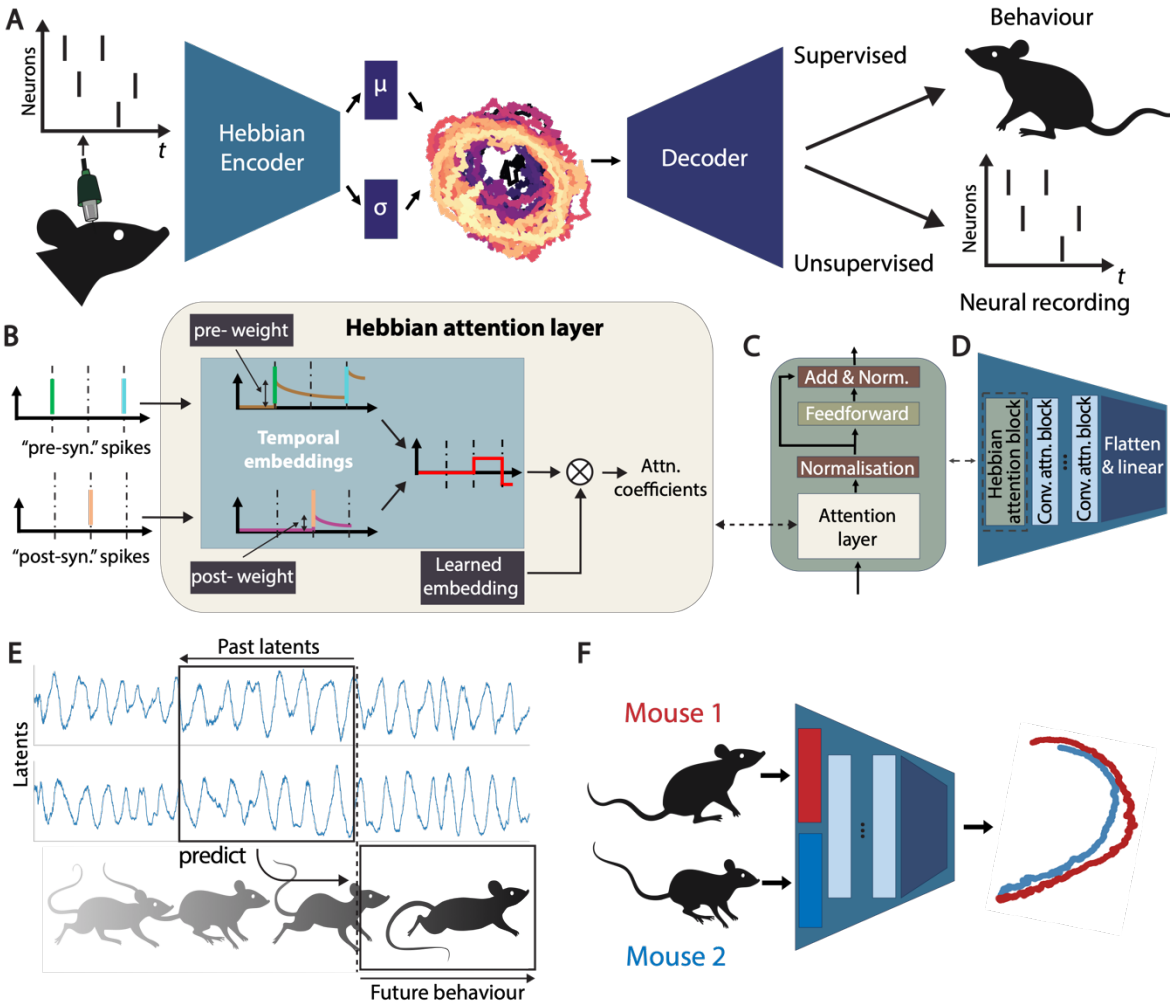

Figure 1 ***Overview of SPARKS.*** *(**A**) Architecture of the proposed variational autoencoder operating directly on spiking signals. Blocks* $\boldsymbol{\mu}(\cdot)$ *and* $\boldsymbol{\sigma}(\cdot)$ *represent the mean and standard deviation of the latents distribution. (**B**) Detailed view of the Hebbian attention layer. (**C**) Schematic of the attention blocks, adapted from [17]. (**D**) SPARKS encoder architecture. (**E**) 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. (**F**) Illustration of training with data from multiple sessions or animals. SPARKS learns a different Hebbian attention layer for each session, while keeping the rest of the encoding network shared.*

dimensional neuronal data, as previous research indicates that most of the variance of the neural activity typically lies within low-dimensional subspaces known as *neural manifolds* [11], [12], [13], and a growing body of evidence suggests that the prism of neural manifolds provides information that cannot be found at the level of single neurons [14],[15]. Early studies of population dynamics relied on linear techniques such as PCA, which ignores the temporal nature of neural dynamics and fine-grained information through binning or averaging [11], [12]. A subsequent machine learning (ML) model demonstrated the benefits of variational autoencoders (VAEs) to generate low-dimensional representations of the data and perform prediction [28], but modern approaches often rely on black-box decoding frameworks such as *Transformer* models, sacrificing interpretability for predictive power [16], [17], [18], [19], [20], [21]. In parallel, contrastive learning has become a successful approach to perform representation learning with temporal data [16], although this approach ignores the specific sequential, asynchronous nature of events in neuronal recordings. Beyond the loss of crucial temporal information, many of these techniques also require extensive pre-processing and manual curation, limiting standardisation and reproducibility [22]. This is an issue as few such methods are capable of finding common representations of the data recorded across sessions and animals [16]. In this work, we propose to combine these two approaches to benefit from the generative, exploratory capabilities of VAEs [23], [24], and the high predictive power of Transformers to improve our understanding of neural data.

Motivated by the growing literature on physics- and biology-inspired AI models, which promises to develop more powerful and explainable AI [25], our contributions are two-fold. Firstly, we integrate Transformers into VAEs and adapt them to the unique nature of neural data using one of the processes behind learning in

biological brains. By designing a *Hebbian attention layer* to operate directly on sparse populations of asynchronous events, we developed a computational framework that is better matched to the structure of neural data than the components of a standard Transformer. Unlike black-box deep learning models, the coefficients learned by this model are interpretable and can be used to gain biological insights about network interactions. Secondly, we introduce a novel criterion inspired by predictive coding, which maximises the probability for the model to predict the correct output for a number of time-steps in the future, given the current latent embeddings. Using this combination, we obtain state-of-the-art predictive power and robust latent representations to analyse neural recordings. Latent embeddings obtained from recordings in different sessions or animals can be directly compared within the same latent space, which will be key towards the effort of building a brain-wide understanding of neural mechanisms.

The proposed model is a Sequential Predictive Autoencoder for the Representation of spiKing Signals (SPARKS). Our approach offers several key advantages: training directly on spiking signals or calcium imaging data with minimal preprocessing; directly interpretable low-dimensional latent embeddings; state-of-the-art predictive power from a small number of neurons; unsupervised exploration of the data; directly interpretable coefficients in the Hebbian attention layer; and robust encodings across repetitions, sessions, animals, and brain areas. Throughout, we showcase SPARKS' capabilities using diverse datasets, highlighting its potential to uncover biological results.

SPARKS: 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 VAE with our Hebbian attention layer and predictive learning rule (Figure 1A-E). Following the seminal work [26], the encoder comprises several attention blocks, each composed of an attention layer (Figure 1B), followed by a fully connected feedforward network with residual connections and batch normalisation (Figure 1C-D). 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 ([26], Figure S 1B). To obtain latent representations of the input data, the output of the last attention block is flattened and projected into the desired latent dimension. Throughout, the decoder is a fully connected feedforward neural network, tasked to predict a desired *reference* signal, albeit it could be chosen to be a recurrent or spiking neural network. 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. Training via unsupervised learning allows the exploration of low dimensional representations of the data produced by the encoder without imposing on the model prior information about task variables. This is particularly interesting when investigating how complex combinations of external and internal parameters are encoded in the data. This analysis can be complemented by supervised training to understand how external parameters recorded during the experiments are encoded in the neural activity.

We train SPARKS through a novel objective which is obtained by extending the definition of causally conditioned distributions to the framework of predictive coding [27], [28]. 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 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 1E). The use of the PCC is inspired by the predictive coding theory, namely, that the encodings produced by the brain's internal generative model constantly aim to predicting future sensory inputs [29]. By forcing the model to learn representations that are predictive of longer-term future dynamics, the proposed objective introduces temporal coherence in the representations produced to capture the true underlying generative process of the data. This aligns with recent discoveries in Transformers, where training models to predict multiple future tokens simultaneously has been shown to improve both performance and training efficiency [30], and we propose predictive learning as a generally applicable and highly effective principle for training temporal models of neural data.

Applying Self-Attention to Neural Data

Transformers are powered by *self-attention* mechanisms [26], which operate in two steps. Firstly, they learn *positional encodings* to make sense of the order of *tokens* in the sequence, before computing relationships between different parts of their input sequence for the model to determine which are relevant. We propose that

Hebbian mechanisms offer an implementation of attention for neural sequences by fusing these two steps via the computation of synaptic strength coefficients based on the relative times of spikes emitted by pre- and post-synaptic neurons. This *Hebbian attention* is a fundamentally spatiotemporal operator: it is sensitive to both which neurons are firing (the spatial component) and when they are firing relative to one another (the temporal component, captured by dynamically evolving eligibility traces) (Figure 1B, Figure S 1A). The Hebbian attention layer implements a recurrent process, which can be embedded into a VAE architecture to benefit from the previously demonstrated capabilities of recurrent VAEs for neural data analysis [31]. This is unlike conventional dot-product attention, which relies on various positional embedding techniques to introduce temporality across tokens presented simultaneously to the model ([26], Figure S 1B).

We sequentially compute attention between neural sequences using 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). While the time constants of decay and the update values for pre- and post-synaptic traces could be measured experimentally [32], they are unknown *a priori*; and therefore, we learn these values for each neuron pair via backpropagation. Our method can also be readily adapted to operate using calcium traces obtained through imaging (Methods). The coefficients obtained via STDP can be likened to the dot-product between keys and queries in the conventional dot-product attention [26] (Figure 1C, Figure S 1B). These values are then projected through a lower-dimensional learned embedding to reduce computational complexity when the number of neurons is high.

As we will demonstrate, the implementation of an attention mechanism via causal STDP enables the model to capture sequential information. This is key to encoding neural data where information lies in the precise order of spikes, such as in the entorhinal-hippocampal system [33], as highlighted in the following section.

Hebbian Attention Captures Sequential Information

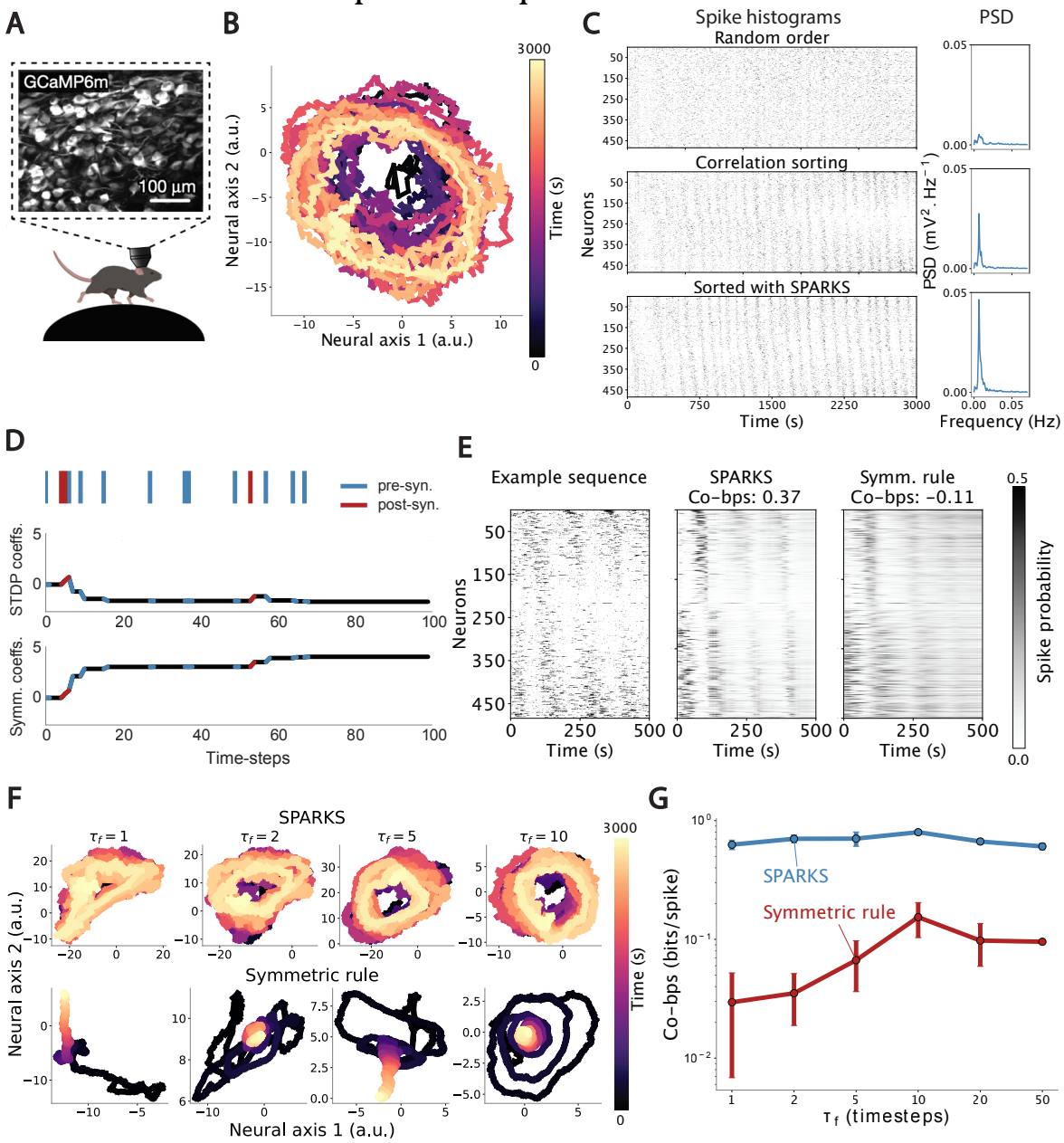

Figure 2 ***SPARKS reveals oscillatory sequences from MEC data.*** *(**A**) Illustration of calcium recording in the MEC of head-fixed mice freely running on a wheel, reproduced from [34]. (**B**) Two-dimensional embeddings obtained from SPARKS for the whole session after unsupervised training. (**C**) Raster plot of thresholded calcium signals of all cells in an example session (bin size = 129 ms, n = 484 cells). Top: random order. Middle: Cells were sorted according to their correlation values with one arbitrary cell, in a descending manner. Bottom: Cells were sorted according to the peak of their correlation with the signal phase recovered with SPARKS. (**D**) Snippet of pre- and post- synaptic firing patterns (top) and corresponding coefficients obtained after application of STDP (middle), and an alternative symmetric rule (bottom) (**E**) Raster plot for left-out data (left) and corresponding predicted spike probabilities after training with SPARKS (centre), and the benchmark symmetric rule (right). (**F**) Evolution of the two-dimensional embeddings obtained from SPARKS (top) and the model with the symmetric Hebbian rule (bottom) for different values of the future prediction window τ_f. (**G**) Evolution of the co-smoothing loss for different values of the future prediction window τ_f.*

We start by illustrating how the latent embeddings obtained with SPARKS can be used to uncover insights from neural data using recordings from the medial entorhinal cortex (MEC, Figure 2A, [34]). It was recently demonstrated that oscillatory sequences are propagated in this structure on the scale of tens of seconds to minutes in passive mice. Uncovering these oscillatory sequences required many analysis steps, each consisting of ad-hoc choices that might result in the masking of relevant information in the neural data by e.g. smoothing [34]. Here, we show that SPARKS can be applied directly to the recordings with minimal pre-processing. Using

a single session that lasted roughly 3,600 seconds, we trained SPARKS via unsupervised learning and found that the 2-dimensional embedding of this data follows a circular trajectory (Figure 2B). We were able to retrieve the phase of the oscillations (Figure S 2A-B, Methods), which allowed us to order neurons according to their correlation to the phase signal to obtain oscillatory sequences (Figure 2C). To assess the quality of the resulting sorting [34], we calculated the height of the largest peak in the power spectral density (PSD) after population decoding of the spike train (Figure 2F, Figure S 2C-D, Methods). The signal recovered after sorting with SPARKS had a peak power of $0.05\ \mathrm{mV}^2.\mathrm{Hz}^{-1}$, outperforming previous methods relying on smooth firing rates ([34], $0.035\ \mathrm{mV}^2.\mathrm{Hz}^{-1}$).

Next, we investigated how our two main contributions, namely the Hebbian attention mechanism and the PCC learning rule, participate in SPARKS' performance. We trained both SPARKS and another VAE-Transformer hybrid model using an alternative symmetric Hebbian attention mechanism (Figure 2D) using the first 3,100 seconds of the recording, and tested them on the remaining 500 seconds. We evaluated the quality of their predictions using a co-smoothing score which calculates how well predictions capture fluctuations in the spiking sequence, as compared to a model predicting average spike rates (Figure 2E, [35], Methods). We found that SPARKS' advantage lies in the use of an asymmetric rule, as the symmetric rule is not capable of producing circular trajectories (Figure 2F, first column). However, importantly, by increasing the value of the prediction window τ_f in the PCC, we were able to recover such oscillations with the symmetric rule. Furthermore, it qualitatively improved the shape of the latent embeddings produced by SPARKS. This also improved the co-smoothing score for both models (Figure 2G), establishing the PCC as a generally relevant rule for learning with temporal models. Interestingly, the effect saturated at $\tau_f = 10$ time-steps, which corresponds to the period of the oscillations. This result further demonstrates that SPARKS' predictive power is closely linked to its capability to model the underlying structure of the population neural manifold. Overall, by incorporating STDP to capture temporal information from neuronal recordings using Hebbian attention, SPARKS can produce latent embeddings that recover information from neuronal populations, while the PCC emerges as a general-purpose tool enabling models to capture temporal information from neural data.

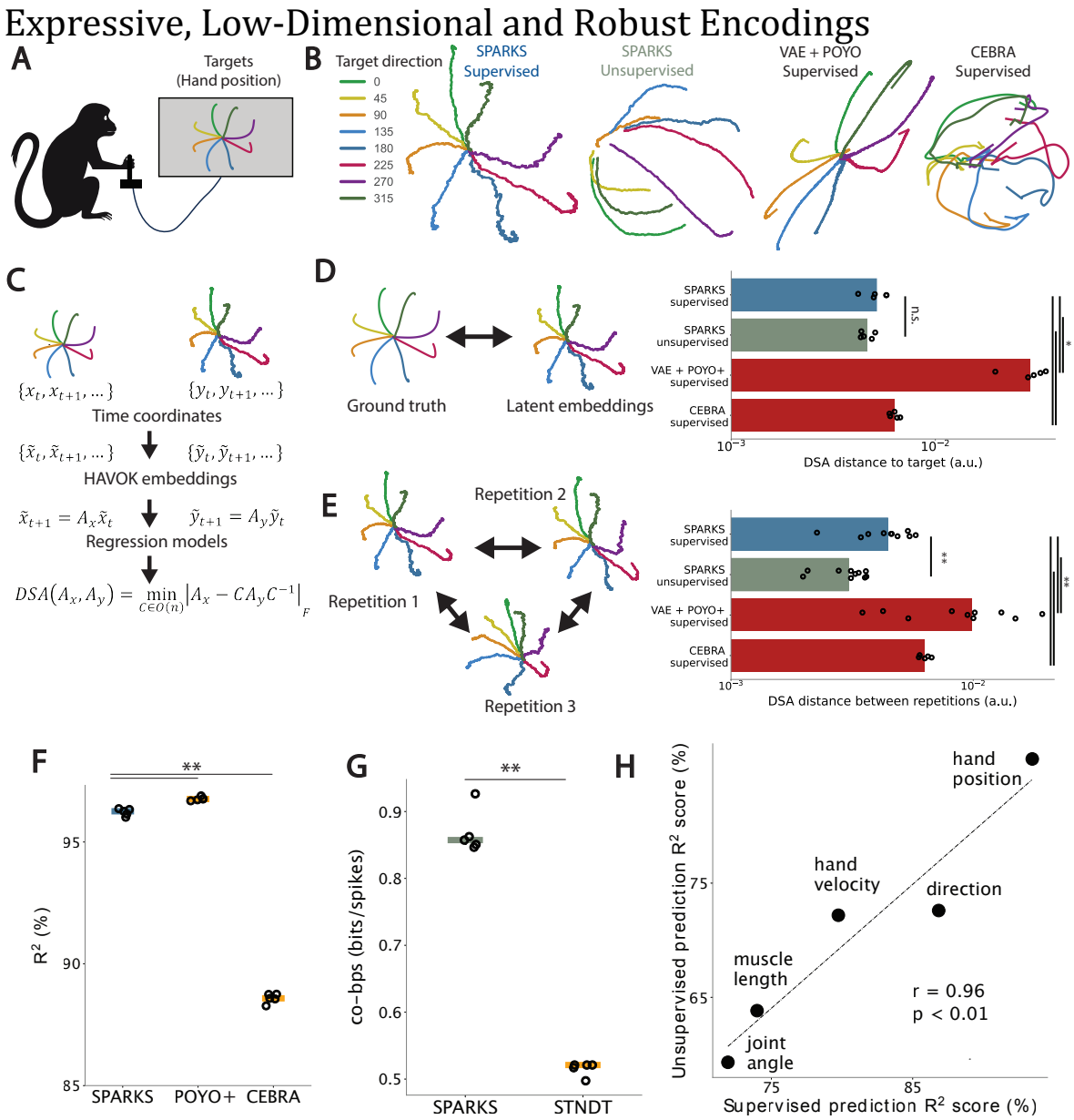

Figure 3 ***Expressive latent embeddings and state-of-the-art prediction from somatosensory cortex data.*** *(**A**) Illustration of the reaching task with average hand positions across all target directions. (**B**) Average latent embeddings obtained through SPARKS for supervised and unsupervised learning, a VAE-Transformer hybrid with POYO+'s attention mechanism [20], and CEBRA . (**C**) Illustration of dynamical systems analysis (DSA). (**D**) Left: Illustration of the comparison between latent embeddings and reference signal. Right: Comparison of the DSA distances between latent embeddings and the task parameter. Each value represents the mean over* 5 *individual runs. (**E**) Left: Illustration of the comparison between latent embeddings across repetitions. Right: Comparison of the DSA distances between the latent embeddings across 5 repetitions after training with random weights initialisation for each model type. (**F**) Comparison of R^2 scores for prediction of hand position. Coloured lines represent the median over 5 runs, and each dot an individual run. (**G**) Comparison of co-smoothing scores for prediction of left-out examples. Orange lines represent the median over 5 runs, and each dot an individual run. (**H**) Unsupervised vs. supervised prediction scores for the same target variables. *: $p < 0.05$, **: $p < 0.01$, n.s.: not significant, two-sided Mann-Whitney U-test. r: Pearson correlation coefficient. p: p-value, t-test.*

To investigate how the latent embeddings obtained via SPARKS can be related to relevant parameters such as behavioural variables, we considered electrophysiology recordings in Brodmann's area 2 of the somatosensory cortex in monkeys performing a centre-out task (Figure 3A). We trained SPARKS on a single recording from the Area2 Bump data set [36], [35]. In this case, the latent embeddings obtained via supervised learning closely match the trajectories of the animal's hand (Figure 3B). Even when training the model via unsupervised learning, latent trajectories are clearly separated across different trial conditions and match the organisation in

space of the target directions. To evaluate the role of the Hebbian attention mechanism, we implemented a model incorporating an instantaneous version of STDP without autoregression (Methods), to find that the latent embeddings it produced were not interpretable (Figure S 3A). However, VAE-Transformer hybrids including conventional scaled dot-product attention with or without autoregression (Methods) both produced interpretable latent embeddings (Figure S 3A), suggesting that the role of this autoregressive process is necessary and specific to Hebbian attention. We also trained a VAE-Transformer hybrid with POYO+'s [19] attention mechanism, which produced smooth embeddings that can be related to the trajectories of the hand (Figure 3B). In contrast, while the embeddings produced by the state-of-the-art model CEBRA [16] are clearly separated by trial direction, they do not visually correspond to the hand-movement trajectories (Figure 3B). Finally, 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 might be necessary to process time-series data. VAE-Transformer hybrids such as SPARKS therefore appear as strong candidates to obtain relevant latent embeddings of neural data.

To quantitatively discriminate between the quality of the latent embeddings produced by different models, we adopted the perspective that the brain implements a non-linear dynamical system, which we aim to approximate in the low-dimensional latent representations produced. In particular, we relied on dynamical systems analysis (DSA, [2]), a method based on Koopman operator theory [37] to compare the vector fields of dynamical systems (Figure 3C). DSA relies on computing high-dimensional approximations of the actual Koopman operators to measure differences between dynamical systems via statistical shape analysis. By computing DSA distances between latent embeddings and trajectories of the hand, we found that SPARKS creates embeddings that are significantly more similar to this reference signal than all benchmarks we implemented (Figure 3D, Figure S 3B). Importantly, although the VAE-Transformer hybrid using POYO+'s [19] attention mechanism produced visually relevant results, these did not recover the dynamics of the reference signal. This demonstrates the advantages of using SPARKS' Hebbian attention over this alternative, which relies on the encoding of individual time-instants. We also used DSA to compute the consistency of latent embeddings across different repetitions of the training procedure with random initialisations. The encodings obtained with SPARKS are found to be significantly more robust than benchmark models, both when obtained through supervised and unsupervised learning (Figure 3E, Figure S 3C).

Another key aspect of SPARKS is its high predictive power, which allowed us to obtain highly accurate reconstruction of task variables (Figure 3F). When training the model via supervised learning, predictive power was quantified as the R^2 score obtained using examples held out for testing. This metric is directly obtained from the decoder's outputs, and we obtain $R^2 = 96.3\%$ for the decoding of hand position, a metric that remains reliable across a large range of hyperparameters (Figure S 3D). SPARKS achieves predictive performance that is competitive with dedicated state-of-the-art decoders (POYO+ [19], $R^2 = 96.8\%$), while simultaneously providing high-quality, dynamically faithful embeddings. It outperformed a VAE-Transformer hybrid equipped with POYO+'s attention mechanism (Figure S 3E), which was incapable of capturing the dynamics of the reference signal, as discussed in the previous paragraph. SPARKS also surpassed the state-of-the-art model for latent representations [16] (R^2 = 88.6%), as well as all other benchmark models we implemented (Figure S 3E). When training SPARKS via unsupervised learning, we compared the reconstructions to those obtained with the state-of-the-art Transformer model for neural data prediction STNDT [21], which we found to perform worse than SPARKS, as measured by co-smoothing. Our results reveal there is no link between predictive accuracy and dynamical fidelity: while models optimised for decoding can achieve high predictive scores, we find that they fail to capture the underlying dynamics of the neural manifold. SPARKS is unique in its ability to achieve both high predictive accuracy and high dynamical fidelity.

When training SPARKS via unsupervised learning, we decoded hand positions from the latent embeddings via an auxiliary decoder, such as a k-nearest neighbours (KNN) model. Remarkably, we found that our unsupervised model achieves similar accuracy to the supervised version of the recent benchmark [16], achieving ($p = 0.16$, Figure S 3F). We then 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 a 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 $r = 0.96$, $p < 0.01$, 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 sensory, motor and state dependent signals.

Taken together, these results demonstrate how SPARKS produces meaningful and robust low-dimensional representations of neural data, providing state-of-the-art performance for supervised and unsupervised learning as well as generation of latent embeddings with only tens of neurons.

Reconstructing Visual Stimuli Across Sessions and Animals

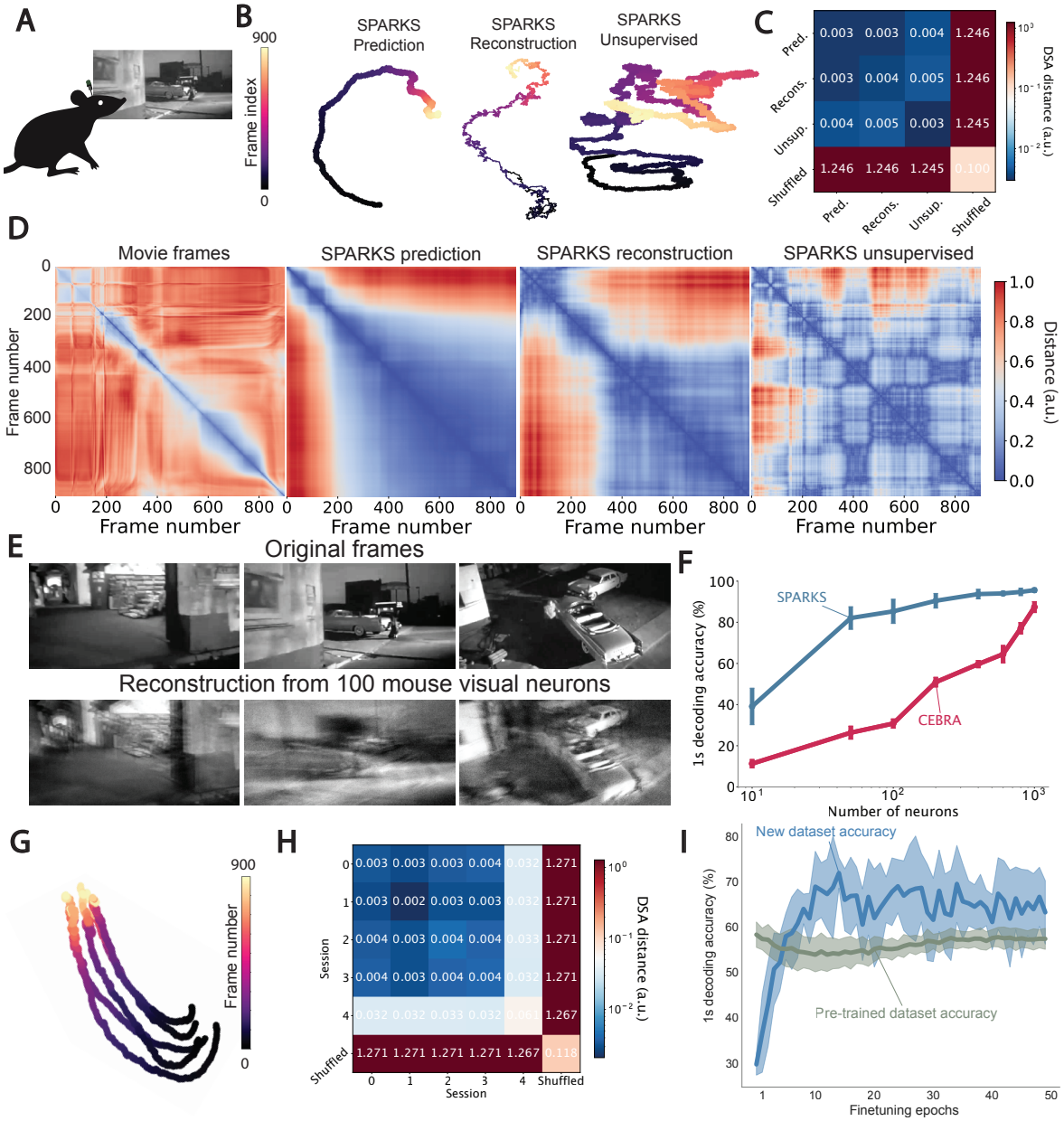

Figure 4 ***Frame reconstruction and prediction from primary visual cortex recordings.*** *(**A**) Schematic of Neuropixels recordings in passive mice viewing a naturalistic film. (**B**) Representative examples of two-dimensional embeddings obtained from the neuronal responses to natural movies. From left to right: SPARKS trained for frame prediction, SPARKS trained for frame reconstruction, SPARKS trained via unsupervised learning. (**C**) DSA distances between latent embeddings obtained for all modalities, average across 5 repetitions. Shuffled: latent representations obtained via unsupervised learning, randomly shuffled across the time dimension. (**D**) 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**). (**E**) Example frames from the naturalistic movie and corresponding reconstructions using* 100 *neurons recorded from mouse primary visual cortex. (**F**) 1s decoding accuracy as a function of number of neurons for prediction within block (yellow). Within block decoding accuracy is shown for [16] (red). Error bars represent standard deviations over five repetitions with random sampling of individual units from all recordings. (**G**) Two dimensional embeddings corresponding to different sessions obtained from the multi-session model. (**H**) DSA distances between latent embeddings obtained for all sessions, average across 5 repetitions. Shuffled: latent representations from session 0, randomly shuffled across the time dimension. (**I**) 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.*

Next, we evaluated SPARKS' capability to study responses to higher-dimensional stimuli such as natural movies, 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 [38], using passive recordings from mice viewing a 30 second natural movie (Figure 4A). In each session, the movie was presented 20 times in two blocks of 10 repetitions, separated by approximately 90 minutes. We tackled the tasks of predicting frame numbers; of reconstructing individual movie frames from neural activity; and of predicting neural activity (unsupervised learning). To do so, we constructed a *pseudo-mouse* by aggregating 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 in block one, leaving the last one out for testing.

Firstly, we evaluated the latent representations obtained via prediction, reconstruction and unsupervised learning. Although the neural trajectories obtained across modalities looked different by visual inspection (Figure 4B), their dynamics were closely related, as measured by DSA (Figure 4C). To understand how these embeddings relate to visual information, we computed normalised distances between movie frames, and normalised pairwise distances between the frame embeddings produced by each model (Figure 4D, Methods). We found that depending on the information contained in the reference signal used for training, different patterns emerged from the latent embeddings. Starting with frame ID prediction, we found that frames that are distant in time were set apart in the latent space, with clusters emerging that correspond to global changes in the visual scene. For frames reconstruction, we found that a finer structure emerged in those clusters with patterns within scenes, relating to smaller changes in the visual information presented to the mice. This effect was most striking in the latent embeddings captured via unsupervised learning, which highlighted the richness of the information stored in the neural population. These changes could not be found simply by looking at pairwise correlations between movie frames, but stemmed from fine-grained changes in the visual scene. This illustrates the advantages of accounting for population dynamics compared to analysing single neuron responses.

To demonstrate the predictive capabilities of SPARKS in this dataset, we reconstructed frames from neural data (Figure 4E). The model is seen to capture high-level scene context, as well as small, fast-moving details (Suppl. Movie 1). For quantitative evaluation, we tested our model on the task of predicting frame IDs (i.e., indices ranging from 0-900, 30s movie, 30 fps, Figure 4F). 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. 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 2 times fewer neurons (SPARKS, 93.7% accuracy, $N = 400$, $p < 0.01$), suggesting that SPARKS is able to capture key information from the neural data within a small neuronal population. Neither a VAE-Transformer hybrid with conventional dot-product attention or a model implementing the symmetrical Hebbian attention rule performed favourably against SPARKS (Figure S 4A). Additionally, we performed frame prediction using spiking data obtained from pre-processed calcium recordings sessions from the same dataset [39]. We found that SPARKS still performs similarly when decoding information from calcium dynamics (Figure S 4B), establishing our algorithm as the current state-of-the-art for prediction from both types of neural recordings.

Generalising neural signal analysis across sessions and animals remains a central goal for any robust and comprehensive modelling approach. Despite recent methodological advances [5], [40], achieving stable chronic electrophysiological recordings in mice continues to be highly challenging due to technical limitations, including tissue inflammation and gradual electrode displacement. Consequently, many studies rely on acute recordings. Combining data across such recordings is standard practice, particularly when the brain region of interest permits the recording of only a limited number of neurons per session [38], [41]. However, this approach introduces additional complexity. It can obscure meaningful variability in behaviour and internal state across trials and sessions, even when experimental conditions and stimulation protocols are tightly controlled [42], [43]. Moreover, such pooling is often impractical, especially in behavioural tasks where trial lengths vary due to differences in response times across animals, and where validating insights across multiple recordings is further complicated by the constraints noted above. To illustrate how SPARKS can be used to overcome these issues, we first separately trained models for frame prediction using all the neurons from 5 randomly selected sessions, and found a high variance in the predictive accuracy across sessions (Figure S 4C). This effect is not explained by the number of neurons in each recording (Figure S 4D, small positive correlation $r = 0.33$, $p > 0.1$), reinforcing the idea that not all recordings from the same brain region are equally informative. Interestingly, despite the high variance in accuracy across sessions, the embeddings produced by the model remain highly consistent, as measured by DSA (Figure S 4F). To directly compare neural dynamics across sessions, SPARKS can be employed to encode data from several sessions within the same latent space. When applying this procedure to the same 5 sessions from the Allen visual

dataset, the resulting embeddings from each session are closely related, both when measured by the Euclidean distance in the latent space (Figure 4G), and as measured by DSA (Figure 4H). Furthermore, the resulting model maintains a predictive accuracy similar to that obtained when training on single sessions, with a high positive correlation between the two (Figure S 4G, $r = 0.93$, $p < 0.01$). 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. Analysing new sessions from a previously trained model is straightforward via addition of a new Hebbian attention layer and finetuning (Figure 4I). We find that the model reaches its final accuracy on the new session within ~ 10 finetuning epochs, with no loss of accuracy when tested on the sessions it was pre-trained on. Overall, SPARKS offers 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.

Isolating the Neural Code of Visual Information

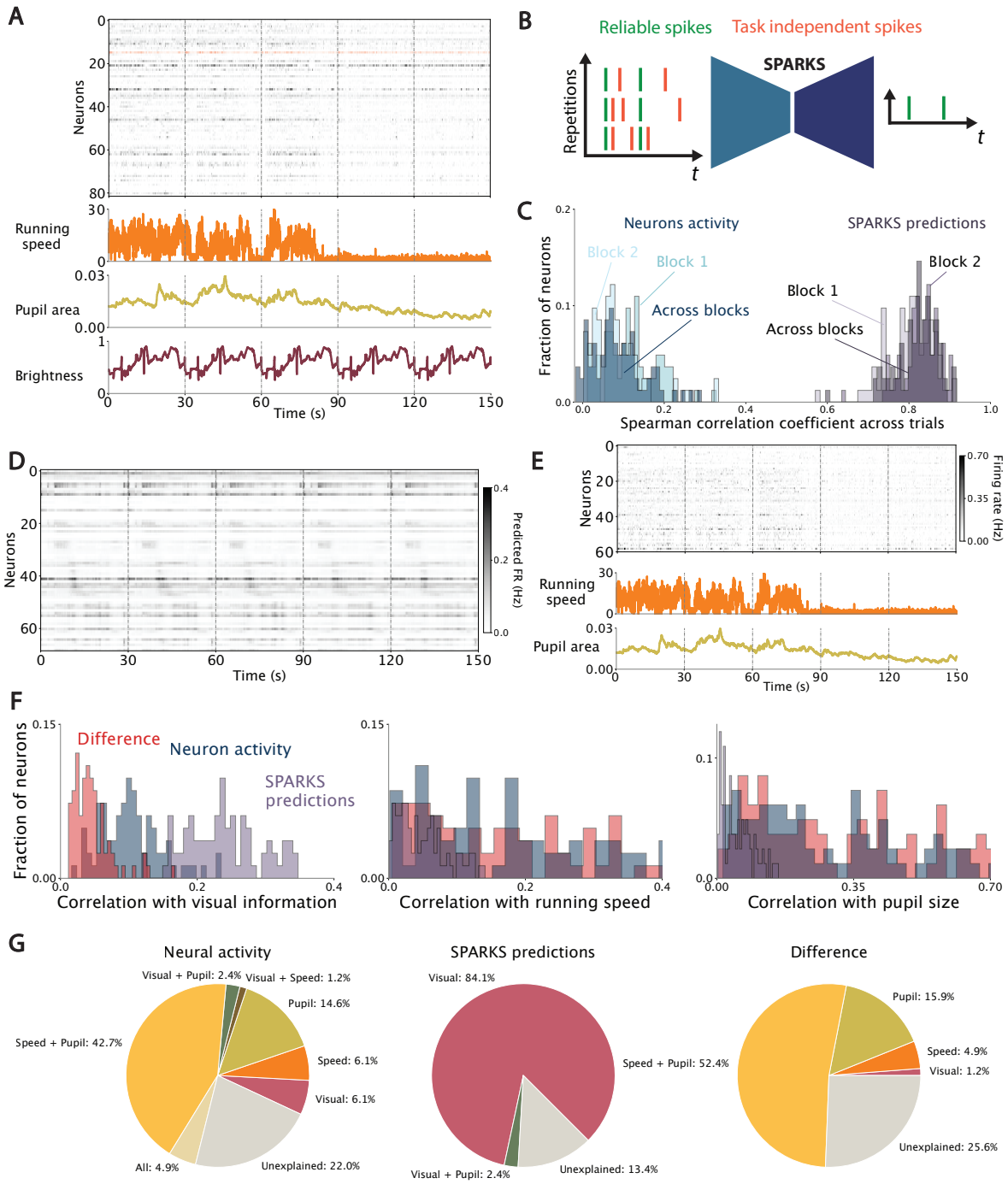

Figure 5 ***Isolating stimulus-driven and behaviour-related neural variance using SPARKS.*** *(**A**) Representative activity of a population of neurons from the primary visual cortex of a passive mouse during presentation of a natural film; running speed, pupil area and average brightness of the 50 pixels most correlated to the neuron whose responses are highlighted in red on the top panel are shown below. Vertical lines correspond to the start of a presentation of the natural movie. (**B**) Illustration of SPARKS as a denoising autoencoder: across repetitions, neural responses present similar patterns that can be learned by the model. (**C**) Spearman correlation of the neurons' responses across trial within each block, and across blocks. (**D**) Raster plot of SPARKS' predicted spiking probabilities, for neurons in which predictions are significantly correlated to pixels brightness. Vertical lines correspond to the start of a presentation of the natural movie. (**E**) Top: Raster plot of the difference between neuronal activity and SPARKS' predictions, for neurons in which the difference is significantly correlated to running, pupil area or both. Middle: running speed. Bottom pupil area for the same animal. Vertical lines correspond to the start of a presentation of the natural movie. (**F**) Distribution of correlation coefficients for raw neuronal activity (blue), SPARKS predictions (purple), and for the difference between the two (pink) with pixels brightness (left), running speed (centre), and pupil size (right). (**G**) Pie charts representing the fractions of neuron correlated to measured parameters, based on their activity (left), SPARKS' prediction (centre), and the difference*

between the two (right).

In the previous section, we focused on forming predictions based on a single block of presentations of the same natural movie, although the movie was presented 20 times in two blocks of 10 repetitions, separated by approximately 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 the high level of variability across trials, due to changes in neural state, behaviour and a phenomenon known as *neural drift* (Figure 5A, [44], [45], [46], [47]). Isolating these effects from reliable visual information encoded in the population is an open challenge in neuroscience research.

To assess whether SPARKS can segregate the source of neuronal variability, we tackled the more challenging task of prediction *across* both blocks and tested our model on the second set of 10 presentations after training on all examples from the first block. Despite the variability of the neural data over time (Figure 5A), we found that SPARKS maintains remarkably high accuracy in this task (Figure S 5A), and the latent embeddings obtained from both blocks are closely matched (Figure S 5B-C). Therefore, by finding the joint structure of the neural manifolds across repetitions of the same visual stimulus, SPARKS is capable of ignoring neural variability to only extract the common features from the neural data, which we postulate to be visual information in this task. To investigate the properties of neuronal variability in these recordings, we tasked SPARKS to learn common features of the neural responses across the two blocks by employing it as a *denoising autoencoder* [48]. Specifically, we trained our model by using as reference signal the neural sequence from another randomly selected movie presentation (Methods), to force it to learn common embeddings across all trials and recover stable visual responses (Figure 5B). We trained the model over 1,000 epochs by randomly selecting the input and reference neural sequences from either block at each epoch. After training, we found that SPARKS had learned common spiking patterns across all presentations of the movie, as demonstrated by the very high level of correlation between the predicted spiking probabilities for all neurons within, but also across blocks (Figure 5C).

As expected, SPARKS' predictions strongly correlated to the brightness of pixels on the screen, which we use here as a proxy for visual information, creating repetitive spiking patterns across repetitions of the movie (Figure 5D). By looking at the difference between actual and predicted spikes, we were able to examine the variable part of the neural responses. We found that the residual signal was often correlated to the running speed of the animal and pupil size (Figure 5E and Figure 5F centre and right columns), but rarely to the visual information (Figure 5F, left column). Although a small (6.1%) fraction of neural responses were correlated to visual information only, we found that the largest group (42.7%) of neurons were correlated to a combination of speed and pupil size, with many cells (22%) presenting no correlation to a parameter recorded in this dataset (Figure 5G). Remarkably, after denoising, the activity of 84.1% of the neurons was strongly correlated with visual content, revealing visual responses that were initially masked by behavioural variance. The remaining activity was correlated to pupil size or unexplained These results demonstrate SPARKS' capability to non-linearly separate the responses of neurons due to various relevant parameters, offering a solution to the key problem of decomposing their variance into sensory responses and behavioural or state-related modulations.

Uncovering the Functional Hierarchy of Mouse Visual Cortical Areas

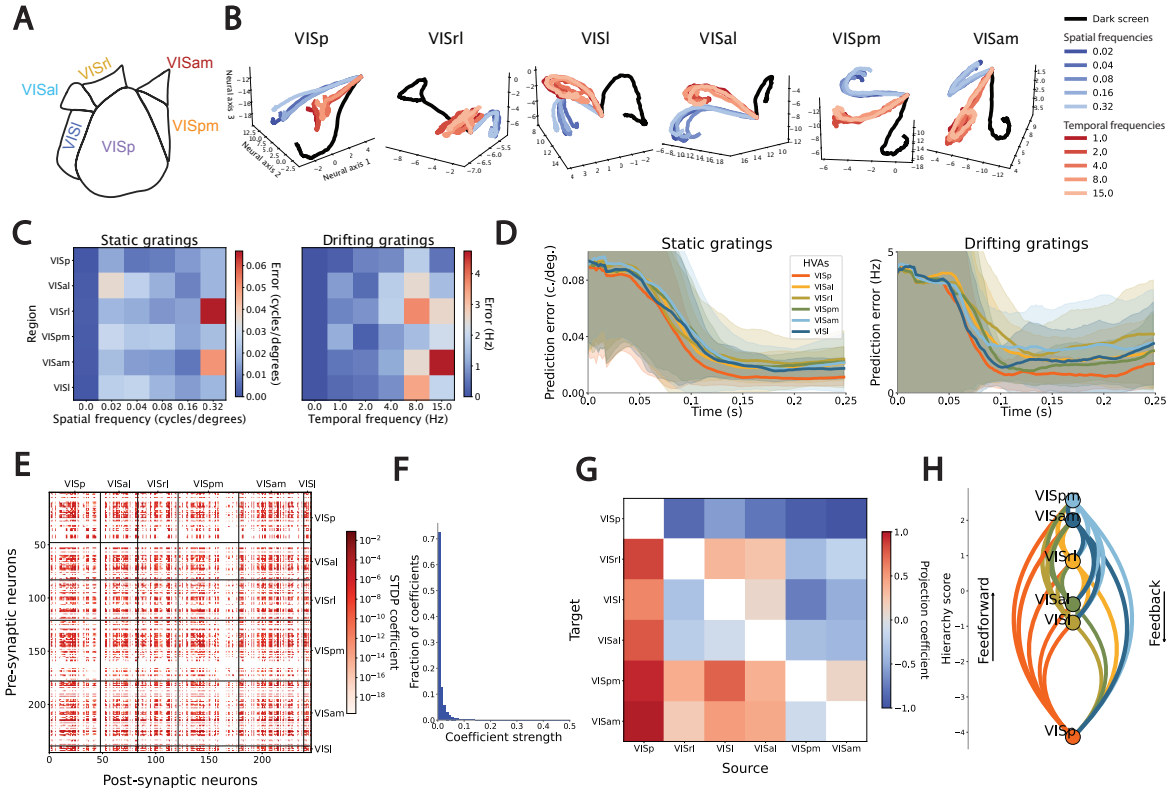

Figure 6 ***SPARKS recovers the hierarchy of the visual cortex.*** *(**A**) Anatomical organisation of mouse visual cortical areas. (**B**) Neural latent representations of static gratings of different spatial frequencies and drifting gratings of different temporal frequencies for each visual area. Responses to a dark screen are shown for comparison. (**C**) Average final spatial and temporal frequency prediction errors for each visual area. (**D**) Evolution of the average spatial and temporal frequency prediction errors over time for each visual cortical area. Lines represent average over 5 individual runs. (**E**) Example STDP coefficients computed as part of the Hebbian attention layer for one session. (**F**) Distribution of all STDP coefficients computed as part of the Hebbian attention layer for 25 sessions. (**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.*

One of the benefits of biologically-inspired models is their increased interpretability, as compared to black-box deep learning models. In SPARKS, the Hebbian attention layer can be used to reveal biological insights in large neuronal networks. To illustrate this, we studied multi-region interactions from Neuropixels recordings obtained from the Allen visual coding dataset acquired from the mouse primary visual cortex (VISp) and five higher-order visual cortical areas (HVAs): latero-medial (LM), antero-lateral (AL), rostro-lateral (RL), postero-medial (PM) and antero-medial (AM) (Figure 6A) [49]. We trained SPARKS using 700 neurons recorded in each HVA from awake, head-fixed mice passively presented with static gratings with various spatial frequencies to predict the spatial and temporal frequencies of the grating presented to the animals at each time-step with a temporal resolution of 1 millisecond.

The time course of responses in the visual cortex of rodents and primates has been linked to frequency preferences, consistent with the coarse-to-fine theory of visual processing [50], [51], [52]. We reasoned that the temporal encoding of spatial and temporal frequencies across HVAs would be informative of the hierarchical organization of the visual system. As seen in Figure 6B, the model trained using data from all HVAs is able to clearly segregate different trajectories for both static and drifting gratings and for all HVAs. SPARKS attains high accuracy for the prediction of spatial and temporal frequencies, with final average errors of 0.018 cycles/degree or 1.2 Hz for static and drifting gratings respectively (Figure 6C). For neural responses from each area, the model starts discriminating between frequencies approximately 50ms from stimulus presentation, but presents different dynamics, with VISp reaching the highest accuracy and being the fastest to achieve its final accuracy (Figure 6D). We notably 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 areal responses: for instance, VISal achieves the lowest error of all areas

only for a medium (0.16 cycles/degrees) spatial frequency. There are also large differences in how quickly gratings are encoded across areas and spatial frequencies; for example, area VISam provides a very fast but poor estimation of high temporal frequencies, while being more accurate but slower for lower temporal frequencies (Figure S 6A).

To uncover interactions between the different HVAs during bottom-up visual stimulation, we analysed the STDP coefficients learned by our Hebbian attention layer after training models via unsupervised learning. To limit the formation of possibly spurious correlations when collating neural responses from different sessions, we trained SPARKS using separate Hebbian attention heads for each of 25 recordings, totalling 5857 neurons (Figure S 6B), and analysed the distribution of the STDP coefficients calculated by the model. We found that they follow a similar distribution to those observed in experimental brain connectivity data, with most areas forming no connections, and presented a heavy tailed distribution (Figure 6F). We computed projections scores using the proportion of non-zero coefficients from each area to all other areas (Methods, Figure S 6C). Similarly to previous works [49], [54], we found that VISp is the area with the most outgoing projections, while areas VISam and VISpm receive the most inputs from other regions (Figure 6G). 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) closely matched the anatomical and functional organization of mouse cortical visual areas previously described [49], [53]. This result can be understood by considering that the STDP coefficients inherently capture directed, temporal relationships between neurons. Therefore, they can be interpreted as a measure of an information flow, which serves as a proxy for the actual functional information flow in the biological circuit, demonstrating the interpretability of SPARKS' Hebbian attention layer to reveal large-scale network interactions.

Discussion

Motivated by the challenges arising from the collection of large-scale neuronal recordings and the surge of generative machine learning techniques, we developed SPARKS, a model combining the benefits of variational autoencoders and Transformers to obtain low-dimensional representations of neural data with very high predictive accuracy. Unlike traditional methods that rely on heavily processed firing rates, SPARKS operates at the single-spike level, capturing the temporal richness of neuronal responses while maintaining interpretability and scalability.

The Hebbian attention layer is the core innovation in SPARKS' architecture, adapting the highly effective attention mechanism of Transformers to process spiking signals. It repurposes the mechanism of STDP into a deep learning framework to offer unique spatio-temporal processing capabilities, as compared to conventional attention mechanisms. Hebbian attention enables our model to process neuronal information more effectively, offering state-of-the-art performance in various tasks and an increase in interpretability. Secondly, we introduce a novel learning objective based on predictive causally conditioned distributions and inspired by predictive coding in the brain. By forcing our model to predict the future evolution of the neural signal, the PCC informs the creation of highly reliable latent embeddings. This is a different approach to the current dominant paradigm of processing time-series data relying on contrastive learning [16], where representations of "positive pairs" are pulled together, while those of "negative pairs" are pushed apart. The PCC objective can be understood as a generative analogue to these contrastive approaches, by constructing an explicit model of the data density relying on the assumption that neural and behavioural states evolve smoothly over short timescales. This objective can therefore be used to train any model for neuronal data, in line with recent work highlighting the benefits for training LLMs to simultaneously predict several tokens [30]. A better understanding of this mechanism and of its links with predictive coding in biological brains will open new perspectives to improve the training of large-scale AI models. Our approach is inspired by the recent successes of physics- and biologically-inspired AI models [25], and demonstrates the relevance of using synergies between AI systems and our biological understanding of intelligence.

As we have highlighted throughout, SPARKS can be used to investigate biological questions using three mechanisms. The first is its capability to discover low-dimensional representations of the high-dimensional neuronal data, or neural manifolds. The neural manifolds approach has been gaining traction since early works in the 2000s demonstrating that "populations are more than the sum of their neurons", that is, the collective firing patterns of population of neurons exhibit properties that cannot be reduced to a simple linear superposition of their firing rates [14],[15]. Recent work notably demonstrated that cortical population activity is resilient to the loss of single neurons, and adapted to maintain its manifolds, demonstrating the relevance of studying neural responses at this level [55]. SPARKS excels at finding neural manifolds by comparing one-to-one the spikes produced by each neuron recorded over time. Understanding the structure of neural manifolds

can shed light on the algorithms the brain implements; a now classical example being the ring attractor topology of head direction in the drosophila [56]. Another approach involves studying the trajectories of neural responses in lower-dimensional spaces, for instance to evaluate how they differ across different sensory conditions by projecting them along different *coding directions.* The main limitation to this approach now lies in our restricted understanding of the mechanisms of generative AI. In this work, we chose to study the latent neural manifolds produced by SPARKS under the perspective of dynamical systems, acknowledging that they cannot reliably be understood under a traditional Euclidean perspective. This offers a theoretical grounding to the analysis of our latent spaces, through which we find that our model is highly reliable, despite the variability inherent to generative models. This angle, that we aim to pursue in future work, notably aligns with recent calls for an algorithmic understanding of generative AI [57], and we propose that biologically-inspired models offer an avenue towards higher model interpretability. In SPARKS, we found that the STDP coefficients obtained within the Hebbian attention layer can be used to uncover the functional structure of large-scale neuronal populations, offering a second approach to analyse data. It comes as an alternative to existing methods such as Granger causality [58], where SPARKS provides an indirect, model-based inference of hierarchy, whereas GC is a direct statistical model of causality.

A final, complementary approach relies on supervised learning to predict relevant parameters such as behavioural outputs or sensory information. SPARKS' state-of-the-art prediction capabilities can find relationships between neural activity and behaviour that previous, less performant models could not capture. In addition, our model's structure opens new possibilities to solve challenging non-linear tasks, such as segregating relevant neural information from behavioural variance. Note that this approach differs from disentangled representation learning [24], [59], [60] in that it does not aim to learn individual subspaces corresponding to distinct, interpretable factors of variation in the data. Rather, it focuses on isolating the stable, stimulus-driven parts of neural responses, with the benefit of not requiring explicit labels for behavioural variables during training. Importantly, since disentangled representation learning relies on VAEs, both approaches could be combined to yield greater benefits. Although predictive power was not our main goal, SPARKS' impressive performance demonstrates the relevance of Hebbian attention to process neuronal data, and places the Hebbian attention layer as a possible building block for the creation of *foundation* models of the brain. These models, comprising many billions of parameters and trained on large quantities of data, are expected to provide novel insights on brain function thanks to their *emergent* properties – that is, the capacity to perform tasks they were not specifically trained on.

Our modelling can also be seen under the perspective of the Bayesian brain hypothesis, by considering SPARKS as a Bayesian encoding-decoding machine [61]. 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, which 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. This approach differs from existing unified encoding-decoding models [23] that rely on purely discriminative masked autoencoders, and instead offers an unsupervised generative model to explore the data.

Despite these advances, important challenges remain. The quadratic complexity of attention limits scalability to the largest imaging datasets, and current implementations cannot yet handle millions of neurons recorded simultaneously. Future work will explore sparse attention mechanisms [62] and the integration of anatomical connectivity to align model structure with the brain's wiring constraints. Additionally, while SPARKS provides compelling insights into latent organization, understanding how these representations map onto circuit mechanisms and behaviour will require targeted experimental validation. Addressing these challenges will not only enhance SPARKS but also inform the development of next-generation brain-inspired AI systems.

In conclusion, SPARKS marks a step toward a unified computational framework for neural data analysis, combining dimensionality reduction, predictive power, and biological interpretability. By integrating Hebbian attention with a predictive coding-inspired objective in a hybrid VAE–Transformer architecture, it matches or surpasses specialized state-of-the-art methods in diverse tasks — from decoding behaviour to revealing large-scale functional organization. As datasets grow in scale and complexity, versatile and scalable tools like SPARKS will be essential for translating neural activity into insights about how the brain represents and predicts the world.

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] 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.

[8] 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.

[9] 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.

[10] N. D. Nguyen *et al.*, 'Cortical reactivations predict future sensory responses', *Nature*, vol. 625, no. 7993, pp. 110–118, Jan. 2024, doi: 10.1038/s41586-023-06810-1.

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

[12] 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.

[13] 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.

[14] J. A. Gallego, 'Neural manifolds: more than the sum of their neurons', *Nat. Rev. Neurosci.*, vol. 26, no. 6, pp. 312–312, Jun. 2025, doi: 10.1038/s41583-025-00919-0.

[15] M. G. Perich, D. Narain, and J. A. Gallego, 'A neural manifold view of the brain', *Nat. Neurosci.*, pp. 1–16, Jul. 2025, doi: 10.1038/s41593-025-02031-z.

[16] S. Schneider, J. H. Lee, and M. W. Mathis, 'Learnable latent embeddings for joint behavioural and neural analysis', *Nature*, vol. 617, no. 7960, pp. 360–368, May 2023, doi: 10.1038/s41586-023-06031-6.

[17] J. Ye and C. Pandarinath, 'Representation learning for neural population activity with Neural Data Transformers', *Neurons Behav. Data Anal. Theory*, vol. 5, no. 3, Aug. 2021, doi: 10.51628/001c.27358.

[18] J. Ye, J. L. Collinger, L. Wehbe, and R. Gaunt, 'Neural Data Transformer 2: Multi-context Pretraining for Neural Spiking Activity', *bioRxiv*, p. 2023.09.18.558113, Sep. 2023, doi: 10.1101/2023.09.18.558113.

[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] M. Azabou, K. X. Pan, V. Arora, I. J. Knight, E. L. Dyer, and B. A. Richards, 'Multi-session, multi-task neural decoding from distinct cell-types and brain regions', presented at the The Thirteenth International Conference on Learning Representations, Oct. 2024. Accessed: Jun. 26, 2025. [Online]. Available: https://openreview.net/forum?id=IuU0wcO0mo

[21] T. Le and E. Shlizerman, 'STNDT: Modeling Neural Population Activity with Spatiotemporal Transformers'.

[22] 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.
[23] Y. Zhang *et al.*, ‘Neural Encoding and Decoding at Scale’, Apr. 20, 2025, *arXiv*: arXiv:2504.08201. doi: 10.48550/arXiv.2504.08201.
[24] 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.
[25] D. A. Selby, M. Sprang, J. Ewald, and S. J. Vollmer, ‘Beyond the black box with biologically informed neural networks’, *Nat. Rev. Genet.*, pp. 1–2, Mar. 2025, doi: 10.1038/s41576-025-00826-1.
[26] A. Vaswani *et al.*, ‘Attention is All you Need’.
[27] G. Kramer, ‘Directed information for channels with feedback’, Doctoral Thesis, ETH Zurich, 1998. doi: 10.3929/ethz-a-001988524.
[28] 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
[29] G. B. Keller and T. D. Mrsic-Flogel, ‘Predictive Processing: A Canonical Cortical Computation’, *Neuron*, vol. 100, no. 2, pp. 424–435, Oct. 2018, doi: 10.1016/j.neuron.2018.10.003.
[30] 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.
[31] C. Pandarinath *et al.*, ‘Inferring single-trial neural population dynamics using sequential auto-encoders’, *Nat. Methods*, vol. 15, no. 10, pp. 805–815, Oct. 2018, doi: 10.1038/s41592-018-0109-9.
[32] 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.
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] S. L. Brunton, M. Budišić, E. Kaiser, and J. N. Kutz, ‘Modern Koopman Theory for Dynamical Systems’, *SIAM Rev.*, vol. 64, no. 2, pp. 229–340, May 2022, doi: 10.1137/21M1401243.
[38] 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.
[39] 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.
[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] 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.
[45] 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.
[46] 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.
[47] S. Sadeh and C. Clopath, ‘Contribution of behavioural variability to representational drift’, eLife. Accessed: Jul. 08, 2024. [Online]. Available: https://elifesciences.org/articles/77907
[48] Y. Bengio, L. Yao, G. Alain, and P. Vincent, ‘Generalized Denoising Auto-Encoders as Generative

Models'.
[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. Noda *et al.*, 'Homeostasis of a representational map in the neocortex', *Nat. Neurosci.*, vol. 28, no. 7, pp. 1533–1545, Jul. 2025, doi: 10.1038/s41593-025-01982-7.
[56] D. E. Angelaki and J. Laurens, 'The head direction cell network: attractor dynamics, integration within the navigation system, and three-dimensional properties', *Curr. Opin. Neurobiol.*, vol. 60, pp. 136–144, Feb. 2020, doi: 10.1016/j.conb.2019.12.002.
[57] O. Eberle, T. A. McGee, H. Giaffar, T. W. Webb, and I. Momennejad, 'Position: We Need An Algorithmic Understanding of Generative AI', presented at the Forty-second International Conference on Machine Learning Position Paper Track, Jun. 2025. Accessed: Jul. 17, 2025. [Online]. Available: https://openreview.net/forum?id=eax2ixyeQL
[58] K. J. Friston, 'Functional and Effective Connectivity: A Review', *Brain Connect.*, vol. 1, no. 1, pp. 13–36, Jan. 2011, doi: 10.1089/brain.2011.0008.
[59] K. J. Miller, M. Eckstein, M. M. Botvinick, and Z. Kurth-Nelson, 'Cognitive Model Discovery via Disentangled RNNs', Jun. 26, 2023, *Cold Spring Harbor Laboratory*. doi: 10.1101/2023.06.23.546250.
[60] D. Yi, S. Musall, A. Churchland, N. Padilla-Coreano, and S. Saxena, 'Disentangled multi-subject and social behavioral representations through a constrained subspace variational autoencoder (CS-VAE)', *eLife*, vol. 12, Jul. 2023, doi: 10.7554/eLife.88602.1.
[61] 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.
[62] M. Zaheer *et al.*, 'Big Bird: Transformers for Longer Sequences', Jan. 08, 2021, *arXiv*: arXiv:2007.14062. doi: 10.48550/arXiv.2007.14062.
[63] 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
[64] 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.
[65] 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.
[66] 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.
[67] D. P. Kingma and M. Welling, 'Auto-Encoding Variational Bayes', Dec. 10, 2022, *arXiv*: arXiv:1312.6114. doi: 10.48550/arXiv.1312.6114.
[68] J. Sjöström and W. Gerstner, 'Spike-timing dependent plasticity', *Scholarpedia*, vol. 5, no. 2, p. 1362, Feb. 2010, doi: 10.4249/scholarpedia.1362.

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

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

(1)

This distribution captures the causal dependence, in the sense of the directed information flow in time, 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 [27].

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, where 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

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

(2)

Note that the evaluation of objective (*2*) is intractable for problems of practical size (see, e.g., [63]) 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 reparameterisation 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

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

(3)

between latent representation $\boldsymbol{z}$ and reference signal $\boldsymbol{r}$ [28], [64]. 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 [65] 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 [63] to lower bound the decoder log-likelihood at each step $t = 1, \dots, T$

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

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 $\boldsymbol{w} = (\boldsymbol{w}^e, \boldsymbol{w}^d)$, that we maximise via stochastic gradient descent (SGD) and the Adam optimiser [66] 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 [67] as

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

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

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

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

Hebbian 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 STD coefficient between sequences i and j. We consider an online STDP kernel, whereby coefficients for each synapse $(i \to j)_{1 \leq i,j \leq N}$ are obtained by maintaining over time a pair of pre- and post-synaptic *eligibility traces* $a_{ij,t}^{\mathrm{pre}}$ and $a_{ij,t}^{\mathrm{post}}$ where $t \in \{1, \dots, T\}$ and N is the number of recorded neurons. Without loss of generality, we assume T to be the (possibly infinite) 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 $\boldsymbol{w}_{ij,t}^{\mathrm{pre}}$ and $\boldsymbol{w}_{ij,t}^{\mathrm{post}}$, and exponentially decay in the absence of a spike by a value determined by scalars τ_{ij}^{pre} and $\tau_{ij}^{\mathrm{post}}$. 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

$$\begin{aligned} a_{ij,t} := \; & a_{ij,t-1} + s_{i,t} \cdot a_{ij,t}^{\mathrm{post}} - s_{j,t} \cdot a_{ij,t}^{\mathrm{pre}} \\ & + \left(1 - a_{ij,t-1}\right) \cdot s_{j,t} \\ & - a_{ij,t-1} \cdot s_{i,t}, \end{aligned}$$

where $s_{i,t} \in \{0, 1\}$ denotes the state of neuron i at time t, and $a_{ij,t}^{\mathrm{pre}}$ and $a_{ij,t}^{\mathrm{post}}$ are pre- and post-synaptic traces. The first line in this equation is an instantaneous approximation of standard online STDP [68], whereas the second and third line correspond to soft bounds that restrict coefficients to be approximately in the range $[0, 1]$. We find this to be necessary to prevent coefficients from diverging for neurons firing at high rates during long recordings. The eligibility traces can be computed as

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

and

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

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 [26].

The Hebbian attention mechanism can further be expanded to account for several versions of the parameter vectors $\mathbf{T}^{\text{pre},(k)} := \left(\tau_{ij}^{\text{pre},(k)}\right)_{ij}$, $\mathbf{T}^{\text{post},(k)}$, $\boldsymbol{W}^{\text{pre},(k)}$, $\boldsymbol{W}^{\text{post},(k)}$, 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 [26], 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

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

Even in this context, the Hebbian attention mechanism remains highly effective for two key reasons. Firstly, because the eligibility trace in our model is fundamentally a decaying memory of a past event, a raw calcium trace can be viewed as a biologically generated eligibility trace. The relative timing of the changes in these traces between neuron pairs provides a robust, albeit temporally smoothed, proxy for the precise spike-timing relationships that SPARKS is built to capture. Furthermore, beyond the signal shape, it is well-established that correlated fluctuations in calcium activity, even on slower timescales, reflect the underlying functional connectivity [8]. Our model's core strength is identifying these inter-neuronal temporal dependencies.

Alternative Self-Attention Mechanisms

Symmetric Hebbian attention

In Figure 2D, we introduce a symmetric Hebbian attention mechanism, for which coefficients are computed as

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

Hebbian attention without autoregression

In Figure S 3A, we introduce a Hebbian attention mechanism without autoregression, for which coefficients are computed as

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

In both of these models, eligibility traces $a_{ij,t}^{\text{pre}}$ and $a_{ij,t}^{\text{post}}$ were computed in the same fashion as for Hebbian

attention. All the other elements of these model are unchanged as compared to SPARKS.

Conventional attention
For the VAE-Transformer hybrid with conventional scaled dot-product attention we implementd, attention coefficients are computed at every time-instant as

$$\text{Attention}(Q, K, V)_t = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

Conventional attention with autoregression
For the VAE-Transformer hybrid with conventional attention and autoregression model we implemented, the inputs to the model were first passed through an autoregressive layer computing an eligibility trace

$$a_{i,t} := a_{i,t-t} \cdot \exp\left(-\frac{dt}{\tau_s}\right) + w_i \cdot dt \cdot s_{i,t},$$

Where values τ_S and w_i are learnable parameters for each neuron i.

Model evaluation

For unsupervised learning, results were evaluated numerically through co-smoothing. This metric is based on the difference between the log-likelihood of a Poisson model with average firing rates $\boldsymbol{\lambda}$ corresponding to model predictions for left-out neurons outputs $\hat{\boldsymbol{y}}$, and a mean model with average value $\bar{\boldsymbol{\lambda}} = \frac{1}{T}\sum \hat{\boldsymbol{y}}_t$ the mean firing rate of the neurons. Results are measured in bits/spike as

$$\text{bits/spike} = \frac{1}{n_{sp}\log 2}\left(\mathcal{L}(\boldsymbol{\lambda}; \hat{\boldsymbol{y}}_{n,t}) - \mathcal{L}(\bar{\boldsymbol{\lambda}}_n; \hat{\boldsymbol{y}}_{n,t})\right),$$

where n_{sp} is the total number of spikes.

Mouse MEC dataset

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 [34], 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.

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.

We also 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 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 2A). The phase signal we hence obtain is shown in Figure S 2B. For sorting with SPARKS, we sorted neurons by ordering them according to their correlation to this phase signal.

We evaluated the quality of each sorting by population decoding of the sorted signals (Figure S 2C) and plotting their PSDs (Figure S 2D). The population decoding $\bar{\boldsymbol{x}}$ of spiking signal $\boldsymbol{x}$ was obtained as

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

Similarly to the oscillatory score proposed in [34], 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 [36]. 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 [35] to train our model. We use 1 ms bin times, and consider recordings from -100 ms to 500 ms after movement onset. Similarly to [35], 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 [38]. 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 $dist\big(f_i, f_j\big) = 1 - corr(f_i, f_j)$ for every pair of frames $\big(f_i, f_j\big)_{1 \le i,j \le 900}$ 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].

Denoising

When using SPARKS for denoising, we use all 20 presentations of the movie $\{\boldsymbol{x}_i\}_{i=1}^{20} \in \{0,1\}^{N \times 5000}$ in both blocks, with N the number of neurons in the desired recording. At every epoch, we train the model via self-supervised learning to minimise the loss

$$\mathcal{L}_{\boldsymbol{w}} = \sum_{i=1}^{20} \mathcal{L}_{\boldsymbol{w}}\big(\boldsymbol{x}_i, \boldsymbol{x}_j\big) = \sum_{i=1}^{20} \mathbb{E}_{p_{\boldsymbol{w}^e}(\boldsymbol{z}||\boldsymbol{x})} \left[\frac{1}{T} \sum_{t=1}^{T} \log q_{\boldsymbol{w}^d}^{(\tau_p, \tau_f)} \left(\boldsymbol{x}_{j,t} \middle| \middle| \boldsymbol{z}_t\right)\right] - \frac{1}{T} \mathrm{KL}\left(p_{\boldsymbol{w}^e}\big(\boldsymbol{z}||\boldsymbol{x}_j\big) \,||p(\boldsymbol{z})\right),$$

where the reference signal $\boldsymbol{r}$ for input sequence $\boldsymbol{x}_i$ obtained during the ith presentation of the movie has been selected as the input sequence $\boldsymbol{x}_j$ obtained during the jth presentation, with $j \neq i$ selected at random.

Functional Hierarchy of Mouse Visual Cortical Areas

We consider static and drifting 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, and temporal frequencies in {0, 1, 2, 4, 8, 15} Hz. We consider all presentations with a fixed orientation of 150 degrees, and all spatial and temporal phases, in addition to dark screens resulting in a total of 872 training examples and 228 test examples. Examples were downsampled with a bin size of 2 ms. When stacking recordings, we used 700 single unit recordings from each visual cortical area (VISp, VISpm, VISam, VISl, VISal and VISrl), which roughly corresponds to the lowest number of neurons collected in a single area across recordings.

We train SPARKS via supervised learning to predict the spatial and temporal frequencies of the gratings at every time-step, we set $\boldsymbol{r}_t$ as a two-dimensional vector comprising the grating's spatial and temporal frequency. For instance, for a static grating with a spatial frequency of 0.02 cycles/degree, we choose $\boldsymbol{r}_t = [0.02, 0]$ for all $t = 1, \dots, 250$,.
Error at time t for frequency f and cortical area i was computed as the L1 distance

$$e_{i,t}^{f} = |f - q_{\boldsymbol{w}^d}(r_{0,t} \mid \boldsymbol{z}_{t-\tau_p}^{t})|$$

for static gratings, and

$$e_{i,t}^{f} = |f - q_{\boldsymbol{w}^d}(r_{1,t} \mid \boldsymbol{z}_{t-\tau_p}^{t})|$$

for drifting gratings.

We trained SPARKS via unsupervised learning by using neurons from each area in 25 separate recordings and computed STDP coefficients $\boldsymbol{A}_T^k$ for every session $k = 1, \dots, 25$. From these coefficients, we computed normalised projections counts from cortical area i to cortical area j for each recording as

$$\mathrm{proj}_{i\to j}^{k} = \frac{1}{N_i N_j} \sum_{\substack{n\in i \\ m\in j}} \mathbf{1}\{\boldsymbol{A}_{nm,T}^{k} \neq 0\},$$

with N_i and N_j the numbers of neurons recorded from HVA i and j respectively, and $1\{\cdot\}$ is the indicator function, which outputs 1 if the condition within the brackets is met, and 0 otherwise. We computed global projection counts (Figure S 6C) as

$$\mathrm{proj}_{i\to j} = \frac{1}{k} \sum_{k} \mathrm{proj}_{i\to j}^{k},$$

Next, we computed directed projection coefficients (Figure 6G) as

$$\mathrm{dir}_{i\to j} = \mathrm{proj}_{i\to j} - \mathrm{proj}_{j\to i}.$$

Finally, we obtained hierarchy scores as

$$\mathrm{hierarchy}_i = \sum_{j} \mathrm{dir}_{i\to j}.$$

Code availability

Code: https://www.github.com/Iacaruso-lab /sparks.

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). S.S. was supported by the Wellcome Trust (award ref. 225412/Z/22/Z). The authors wish to thank O. Simeone, A. Schaefer, S. Tootoonian, M. Kollo and S. Gonzalo-Cogno for their discussions and comments on the manuscript, and members of the neuronal circuits and behaviour laboratory for their comments and suggestions on the figures.

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. A. E.-W. collected the dataset used for conference presentations of this work. F.I. and S.S. revised the manuscript.

Materials & Correspondence

Requests should be addressed to N. Skatchkovsky.

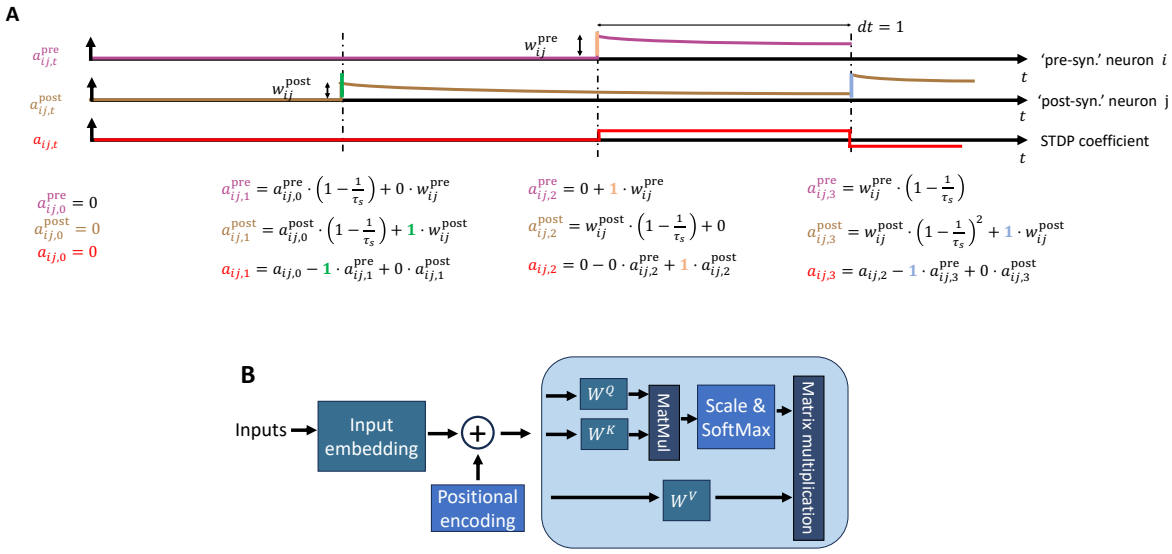

*Figure S 1(**A**) Detailed illustration of the computation of STDP coefficients in the Hebbian attention layer. (**B**) Schematic of the conventional dot-product attention layer, adapted from [26].*

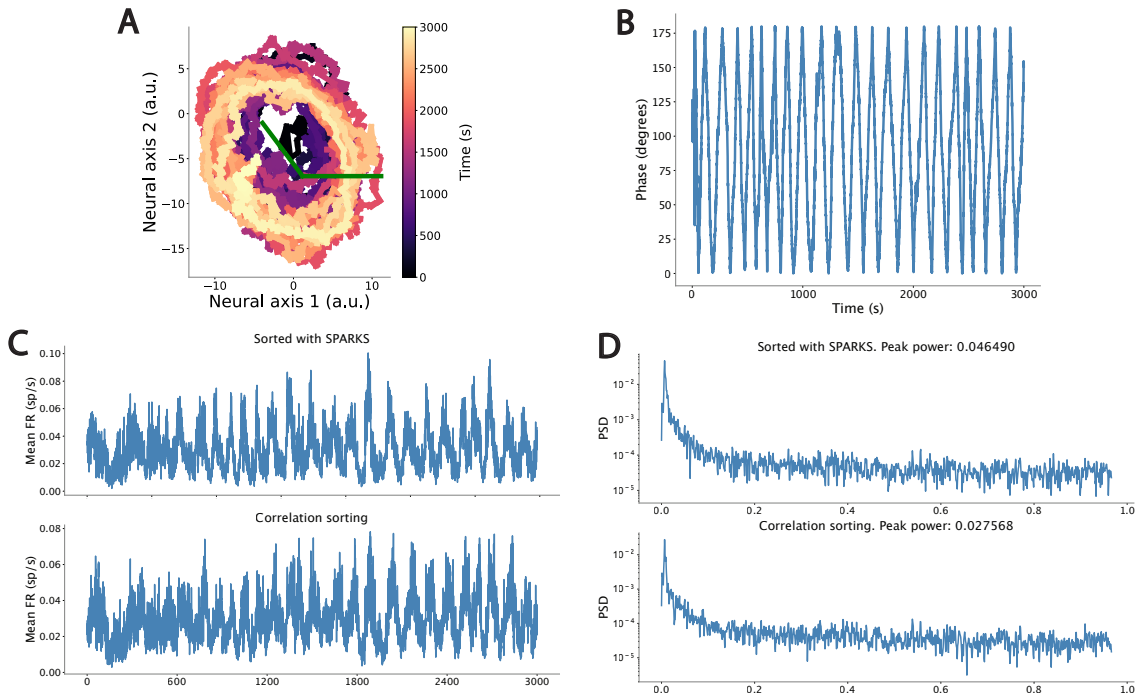

*Figure S 2 (**A**) Illustration of the angle used to recover the phase of the oscillations from the latent embeddings. (**B**) Phase signal recovered from the latent embeddings. (**C**) Oscillatory signals recovered by population decoding of the spike trains after sorting neurons with different methods. (**D**) PSDs corresponding to signals in (**C**).*

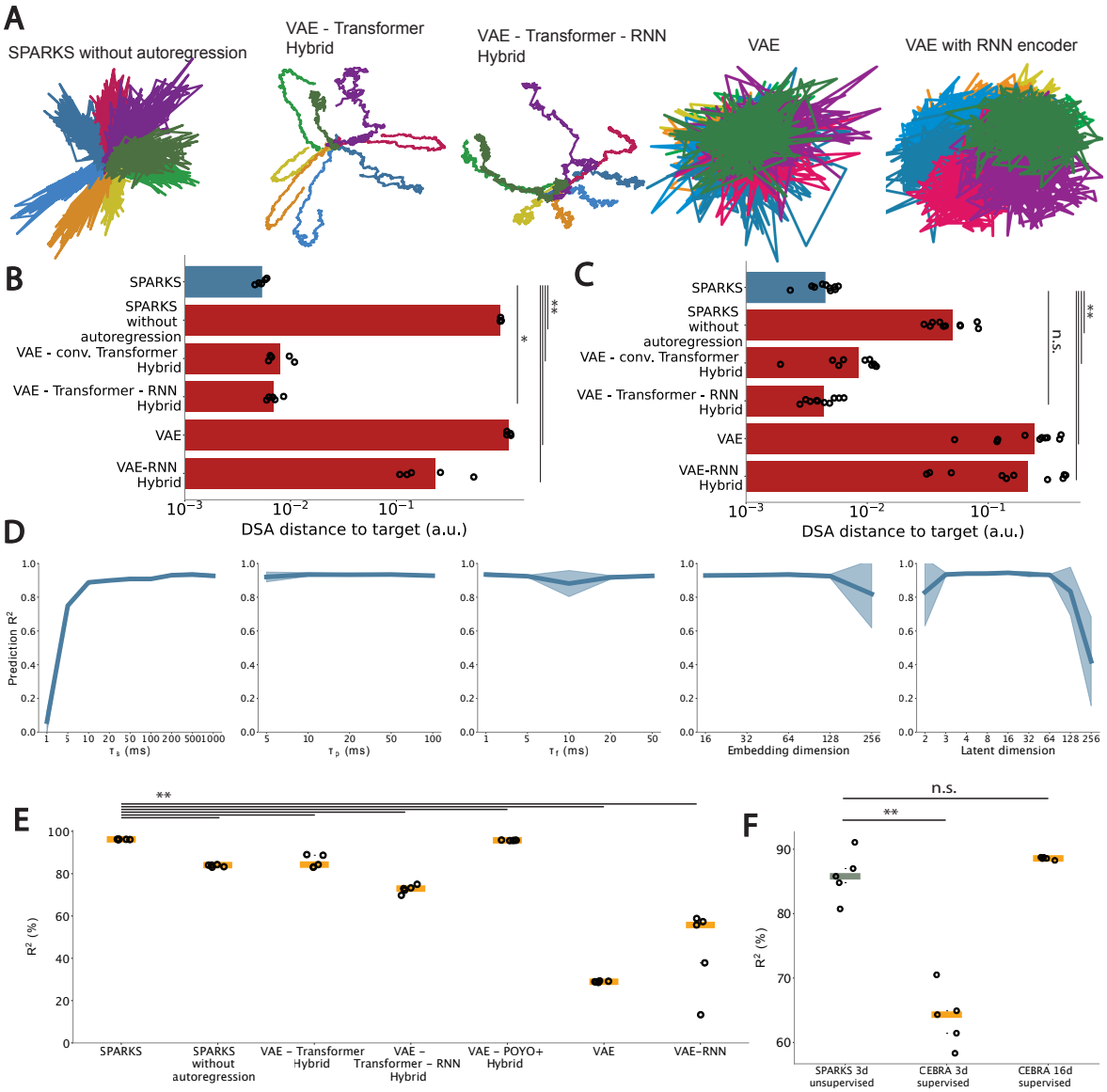

Figure S 3 ***(A)*** *Example latent embeddings for models trained via supervised learning.* ***(B)*** *Comparison of the DSA distances between latent embeddings and the task parameter. Each value represents the mean over* 5 *individual runs, circles represent individual runs.* ***(C)*** *Comparison of the DSA distances between the latent embeddings across 5 repetitions after training with random weights initialisation for each model type. Each value represents the mean over* 5 *individual runs, circles represent individual runs.* ***(D)*** *Prediction* R^2 *scores for different values of the model's hyperparameters. Shaded areas represent standard deviation over* 5 *individual runs.* ***(E)****,* ***(F)*** *Comparison of* R^2 *scores for prediction of hand position. Orange lines represent the median over 5 runs, and each dot an individual run. *:* $p < 0.05$*, **:* $p < 0.01$*, n.s.: not significant, two-sided Mann-Whitney U-test.*

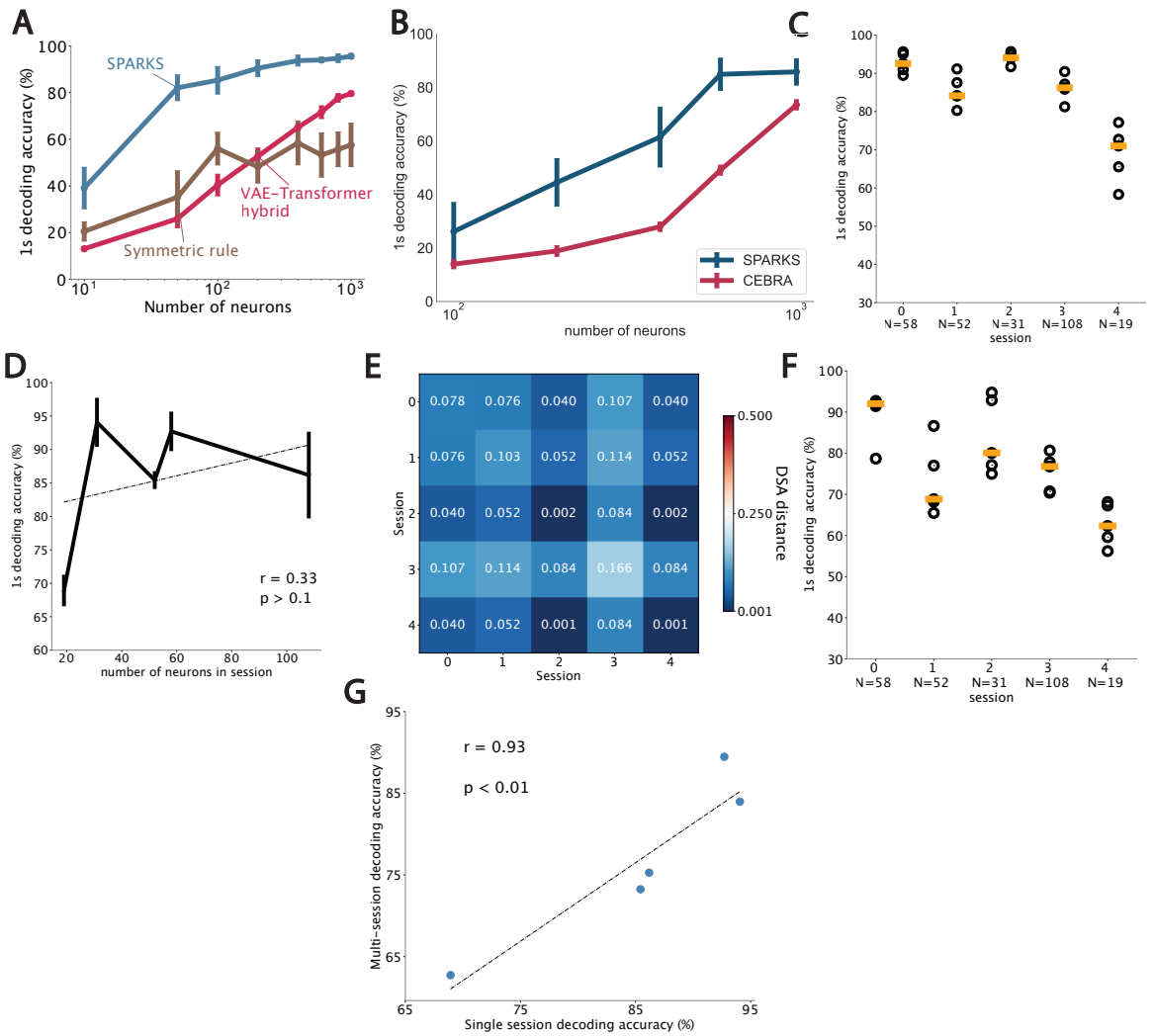

*Figure S 4 (**A**) 1s decoding accuracy as a function of number of randomly sampled neurons for SPARKS, a VAE-Transformer with dot-product attention and a model implementing symmetric Hebbian attention. Error bars represent standard deviation across 5 repetitions. (**B**) 1s decoding accuracy as a function of number of randomly sampled neurons for SPARKS, and CEBRA using deconvoluted calcium imaging data. Error bars represent standard deviation across 5 repetitions (**C**) 1s decoding accuracy for different recording sessions. Orange lines represent the median over 5 runs, and each dot an individual run. (**D**) 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. p: p-value. (**E**) DSA distances between the latent embeddings obtained for different example sessions, averaged over 5 repetitions and 10 movie presentations from block 1. (**F**) 1s decoding accuracy per session for a model trained on 5 different sessions. Orange lines represent the median over 5 runs, and each dot an individual run. (**G**) Correlation between the prediction accuracy for models trained on individual sessions, or multiple sessions simultaneously.*

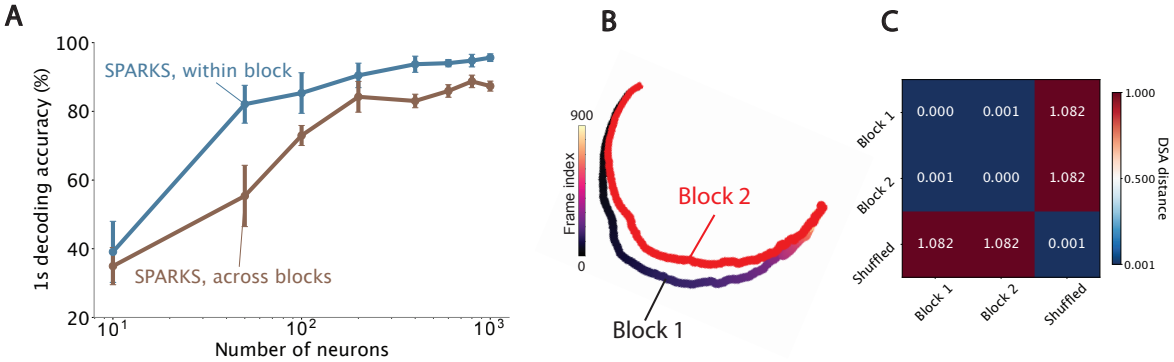

*Figure S 5 (**A**) 1s decoding accuracy as a function of number of neurons for prediction within block (yellow), and across*

*two recording blocks separated by 90 minutes (brown). Error bars represent standard deviations over five repetitions with random sampling of individual units from all recordings. (**B**) Latent embeddings averaged across examples from blocks 1 and 2 (r), for a model trained on block 1 only. (**C**) DSA distances between the embeddings produced for latents obtained from recordings in two different blocks. Shuffled: latent representations from block 1, randomly shuffled across the time dimension. Note the distance between shuffled and shuffled is not meaningful as these examples lack the structure to fit DSA.*

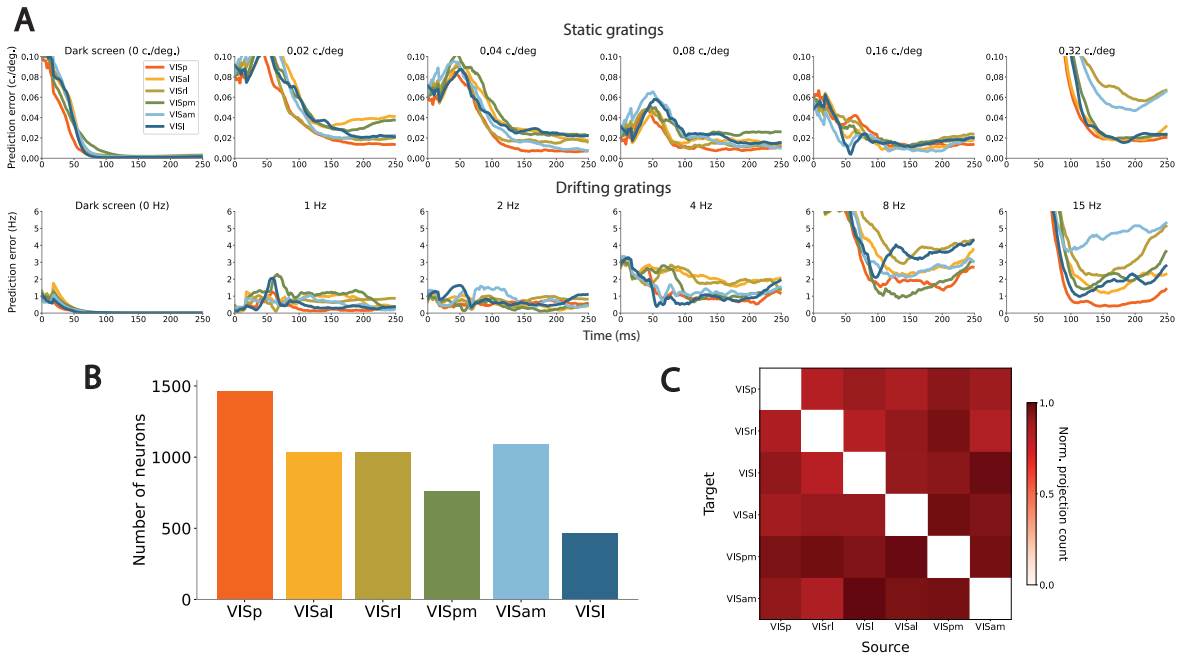

*Figure S 6 (**A**) Evolution of the prediction error over time for each visual cortical area and spatial or temporal frequency. Lines represent average over 5 individual runs. (**B**) Total number of neurons for each HVA. (**C**) Normalised projection calculated from the number of non-zero Hebbian attention coefficients between each HVA.*