

Sensory cortex is optimised for prediction of future input

Yosef Singer¹, Yayoi Teramoto¹, Ben D. B. Willmore¹, Andrew J. King¹, Jan W. H. Schnupp², Nicol S. Harper^{1*}

¹Dept. of Physiology, Anatomy and Genetics (DPAG), Sherrington Building, University of Oxford, Parks Road, Oxford OX1 3PT, UK.

²Dept. of Biomedical Sciences, City University of Hong Kong. 31 To Yuen Street, Kowloon Tong, Hong Kong.

*Corresponding author: nicol.harper@dpag.ox.ac.uk

18 **Neurons in sensory cortex are tuned to diverse features in natural scenes. But what**
19 **determines which features neurons become selective to? Here we explore the idea**
20 **that neuronal selectivity is optimised to represent features in the recent sensory**
21 **past that best predict immediate future inputs. We tested this hypothesis using**
22 **simple feedforward neural networks, which were trained to predict the next few**
23 **video or audio frames in clips of natural scenes. The networks developed receptive**
24 **fields that closely matched those of real cortical neurons in different mammalian**
25 **species, including the oriented spatial tuning of primary visual cortex, the frequency**
26 **selectivity of primary auditory cortex and, most notably, their temporal tuning**
27 **properties. Furthermore, the better a network predicted future inputs the more**
28 **closely its receptive fields resembled those in the brain. This suggests that sensory**
29 **processing is optimised to extract those features with the most capacity to predict**
30 **future input.**

31

32 **Impact statement**

33 **Prediction of future input explains diverse neural tuning properties in sensory**
34 **cortex.**

35

36 Introduction

37 Sensory inputs guide actions, but such actions necessarily lag behind these inputs due to delays
38 caused by sensory transduction, axonal conduction, synaptic transmission, and muscle activation.
39 To strike a cricket ball, for example, one must estimate its future location, not where it is now¹.
40 Prediction has other fundamental theoretical advantages: a system that parsimoniously predicts
41 future inputs from their past, and that generalizes well to new inputs, is likely to contain
42 representations that reflect their underlying causes². This is important because ultimately, we are
43 interested in these causes (e.g. flying cricket balls), not the raw images or sound waves incident on
44 the sensory receptors. Furthermore, much of sensory processing involves discarding irrelevant
45 information, such as that which is not predictive of the future, to arrive at a representation of what is
46 important in the environment for guiding action².

47 Previous theoretical studies have suggested that many neural representations can be
48 understood in terms of efficient coding of natural stimuli in a short time window at or just before
49 the present^{3–6}. Such studies generally built a network model of the brain, which was trained to
50 represent stimuli subject to some set of constraints. One pioneering such study trained a network to
51 efficiently represent static natural images using a sparse, generative model^{5,6}. More recent
52 studies have used related ideas to model the representation of moving (rather than static) images^{7–9}
53 and other sensory stimuli^{10–14}. In contrast, we built a network model that was optimised not for
54 efficient representation of the recent past, but for efficient prediction of the immediate future of the
55 stimulus, which we will refer to as the temporal prediction model. The timescale of prediction
56 considered for our model is in the range of tens to hundreds of milliseconds. Conduction delays to
57 cortex and very fast motor responses are on this timescale^{15–17}.

58 The idea that prediction is an important component of perception dates at least as far back as
59 Helmholtz^{18,19}, although what is meant by prediction and the purpose it serves is quite varied
60 between models incorporating it^{20,21}. With regards to perception and prediction, two contrasting but
61 interrelated frameworks have been distinguished^{20,21}. In the “predictive coding” framework^{22–24},

62 prediction is used to remove statistical redundancy in order to provide an efficient representation of
63 the entire stimulus. Some models of this type use prediction as a term for estimation of the current
64 or a static input (such as images) from latent variables²³, whereas other have also considered the
65 temporal dimension of the input^{25–27}. Sparse coding models can be related to this framework²². In
66 contrast, the “predictive information” framework^{2,21,28,29}, which our approach relates to more
67 closely, involves selective encoding of those features of the stimulus that predict future input. A
68 related idea to predictive information is the encoding of slowly varying features^{8,30–32}, which are
69 one kind of predictive feature. Hence, the predictive coding approach seeks to find a compressed
70 representation of the entire input, whereas the predictive information approach selectivity encodes
71 only predictive features^{20,21}. Our model relates to the predictive information approach in that it is
72 optimized to predict the future from the past, but it has a combination of characteristics, such a non-
73 linear encoder and sparse weight regularization, which have not previously been explored for such
74 an approach.

75 To evaluate the representations produced by these normative theoretical models, they can be
76 optimised for natural stimuli, and the tuning properties of their units compared to the receptive
77 fields of real neurons. A useful and commonly used definition of a neuron’s receptive field (RF) is
78 the stimulus that maximally linearly drives the neuron^{33–38}. In mammalian primary visual cortex
79 (V1), neurons typically respond strongly to oriented edge-like structures moving over a particular
80 retinal location^{39–42}. In mammalian primary auditory cortex (A1), most neurons respond strongly to
81 changes in the amplitude of sounds within a certain frequency range³⁷.

82 The temporal prediction model provides a principled approach to understanding the
83 temporal aspects of RFs. Previous models, based on sparsity or slowness principles, were successful
84 in accounting for many spatial aspects of V1 RF structure^{5–9,43}, and had some success in accounting
85 for spectral aspects of A1 RF structure^{10–12,14}. However, these models do not account well for the
86 temporal structure of V1 or A1 RFs. Notably, for both vision⁴² and audition³⁷, the envelopes of real
87 neuronal RFs tend to be asymmetric in time, with greater sensitivity to very recent inputs compared

88 to inputs further in the past. In contrast, the RFs predicted by previous models^{7,10,11,13,14}
89 typically show symmetrical temporal envelopes, with either approximately flat envelopes over time
90 or a balanced falloff of the envelope over time either side of a peak. They also lack the greater
91 sensitivity to very recent inputs.

92 Here we show using qualitative and quantitative comparisons that, for both V1 and A1 RFs,
93 these shortcomings are largely overcome by the temporal prediction approach. This stands in
94 contrast to previous models, suggesting that neural sensitivity at early levels of the cortical
95 hierarchy may be organised to facilitate a rapid and efficient prediction of what the environment
96 will look like in the next fraction of a second.

97

98 **Results**

99 **The temporal prediction model**

100 To determine what type of sensory RF structures would facilitate predictions of the imminent
101 future, we built a feedforward network model with a single layer of nonlinear hidden units, mapping
102 the inputs to the outputs through weighted connections (Fig. 1). Each hidden unit's output results
103 from a linear mapping (by input weights) from the past input, followed by a monotonic
104 nonlinearity, much like the classic linear-nonlinear model of sensory neurons¹⁰⁻¹². The model then
105 generates a prediction of the future from a linear weighting of the hidden units' outputs. This is
106 consistent with the observation that decoding from the neural response is often well approximated
107 by a linear transformation⁴⁴.

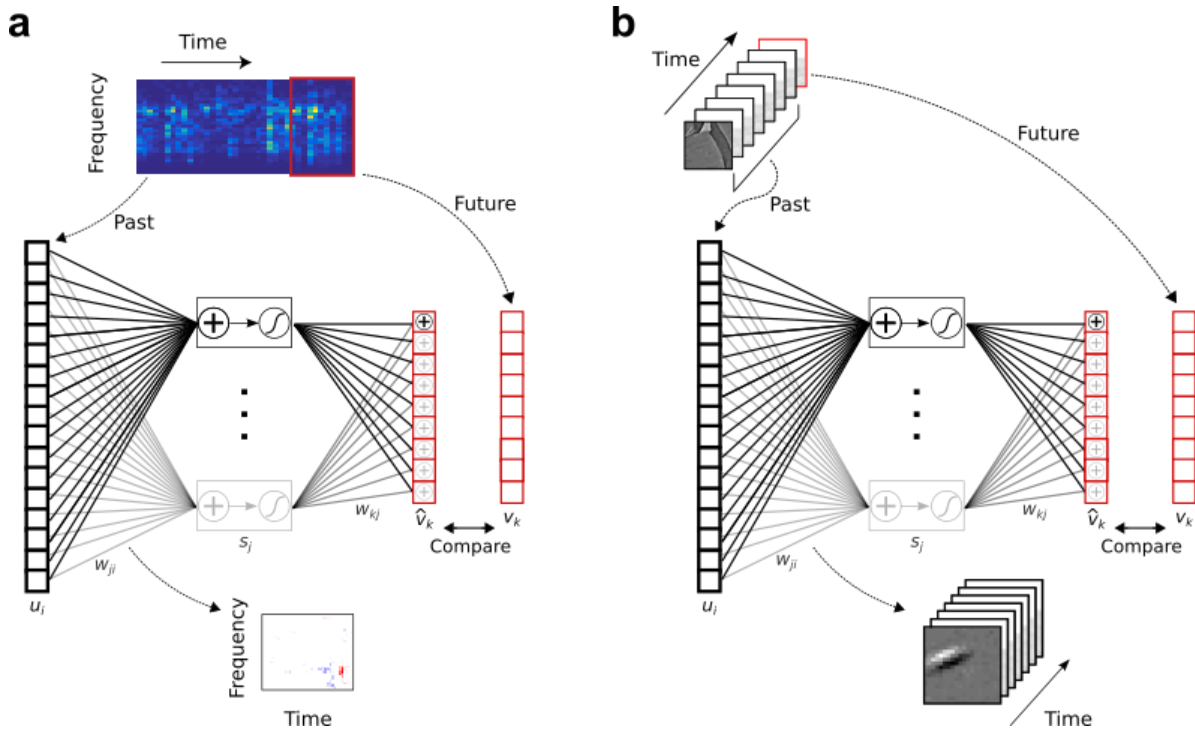


Figure 1 | Temporal prediction model implemented using a feedforward artificial neural network, with the same architecture in both visual and auditory domains. a, Network trained on cochleagram clips (spectral content over time) of natural sounds, aims to predict immediate future time steps of each clip from recent past time steps. **b,** Network trained on movie clips of natural scenes, aims to predict immediate future frame of each clip from recent past frames. u_i , input – the past; w_{ji} , input weights; s_j , hidden unit output; w_{kj} , output weights; \hat{v}_{kn} , output – the predicted future; v_k , target output – the true future. Hidden unit's RF is the w_{ji} between the input and that unit j .

We trained the temporal prediction model on extensive corpora, either of soundscapes or silent movies, modelling A1 (Fig. 1a) or V1 (Fig. 1b) neurons, respectively. In each case, the networks were trained by optimising their synaptic weights to most accurately predict the immediate future of the stimulus from its very recent past. For vision, the inputs were patches of videos of animals moving in natural settings, and we trained the network to predict the pixel values for one movie frame (40 ms) into the future, based on the 7 most recent frames (280 ms). For audition, we trained the network to predict the next three time steps (15 ms) of cochleagrams of natural sounds based on the 40 most recent time steps (200 ms). Cochleagrams resemble spectrograms but are adjusted to approximate the auditory nerve representation of sounds (see Methods).

During training we used sparse, L_1 weight regularisation (see Eqn. 3 in Methods) to constrain the network to predict future stimuli in a parsimonious fashion, forcing the network to use as few weights as possible while maintaining an accurate prediction. This constraint can be viewed as an assumption about the sparse nature of causal dependencies underlying the sensory input, or alternatively as analogous to the energy and space restrictions of neural connectivity. It also prevents our network model from overfitting to its inputs. Note that this sparsity constraint differs from that used in sparse coding models, in that it is applied to the weights rather than the activity of the units, being more like a constraint on the wiring between neurons than a constraint on their firing rates.

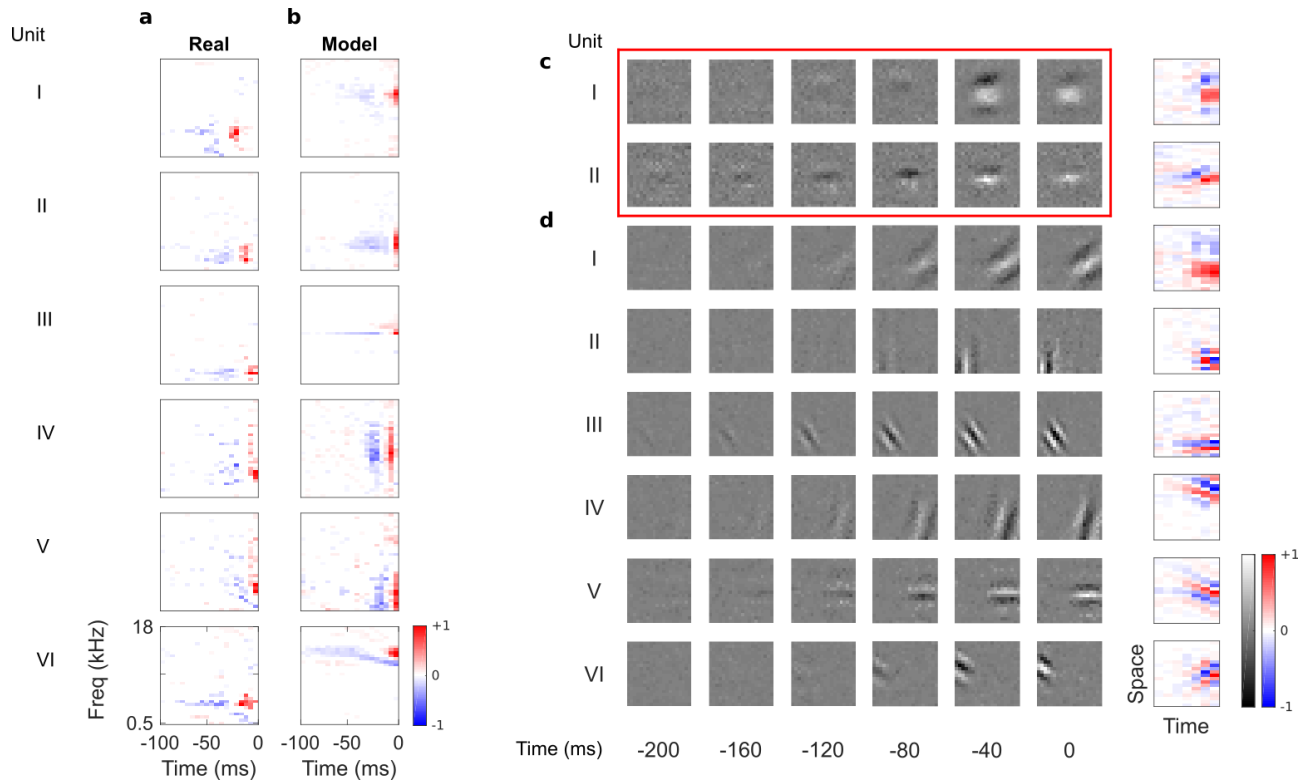
Qualitative assessment of auditory receptive fields

To compare with the model, we recorded responses of 114 auditory neurons (including 76 single units) in A1 and the anterior auditory field of 5 anaesthetised ferrets⁴⁵ and measured their spectrotemporal RFs (see Methods). Ferrets are commonly used for auditory research, because they are readily trained in a range of sound detection, discrimination or localization tasks⁴⁶, the frequency range of their hearing (approximately 40 Hz – 40 kHz⁴⁷) overlaps well with (and extends beyond) the human range, and most of their auditory cortex is not buried in a sulcus and hence easily accessible for electrophysiological or optical measurements.

The A1 RFs we recorded are diverse (Fig. 2a); their frequency tuning can be narrowband, broadband or more complex, sometimes showing flanking inhibition. They may also lack clear order or be selective for the direction of frequency modulation⁴⁸.

In their temporal tuning, A1 RFs tend to weight recent inputs more heavily, with a temporally asymmetric power profile, involving excitation near the present followed by lagging inhibition of a longer duration³⁷. The temporal prediction model RFs (Fig. 2b) are similarly diverse, showing all of the RF types seen *in vivo* (including examples of localised, narrowband, broadband, complex, disordered and directional RFs) and are well matched in scale and form to those measured

155 in A1. This includes having greater power (root mean square) near the present, with brief excitation
 156 followed by longer lagging inhibition, producing an asymmetric power profile. This stands in
 157 contrast to previous attempts to model RFs based on efficient and sparse coding hypotheses, which
 158 either did not capture the diversity of RFs¹², or lacked temporal asymmetry, punctate structure, or
 159 appropriate time scale^{10,11,13,14,48,49}.
 160



161
 162

163 **Figure 2 | Auditory spectrotemporal and visual spatiotemporal RFs of real neurons and**
 164 **temporal prediction model units. a**, Example spectrotemporal RFs of real A1 neurons⁴⁵. Red –
 165 excitation, blue – inhibition. Most recent two time steps (10ms) were removed to account for
 166 conduction delay. **b**, Example spectrotemporal RFs of model units when model is trained to predict
 167 the future of natural sound inputs. Note that the overall sign of a receptive field learned by the
 168 model is arbitrary. Hence, in all figures and analyses we multiplied each model receptive field by -1
 169 where appropriate to obtain receptive fields which all have positive leading excitation (see
 170 Methods). **c**, Example spatiotemporal (I, space-time separable, and II, space-time inseparable)
 171 RFs of real V1 neurons. Left, grayscale: 3D (space-space-time) spatiotemporal RFs showing the
 172 spatial RF at each of the most recent 6 time steps. Most recent time step (40ms) was removed to
 173 account for conduction delay. White – excitation, black – inhibition. Right: corresponding 2D
 174 (space-time) spatiotemporal RFs obtained by summing along the unit's axis of orientation for each
 175 time step. Red – excitation, blue – inhibition. **d**, Example 3D and corresponding 2D spatiotemporal
 176 (I-III, space-time separable, and IV-VI, space-time inseparable) RFs of model units when model is
 177 trained to predict the future of natural visual inputs.
 178

Qualitative assessment of visual receptive fields

By eye, substantial similarities were also apparent when we compared the temporal prediction model's RFs trained using visual inputs (Fig. 1b) with the 3D (space-space-time) and 2D (space-time) spatiotemporal RFs of real V1 simple cells, which were obtained from Ohzawa et al⁵⁰. Simple cells³⁹ have stereotyped RFs containing parallel, spatially localised excitatory and inhibitory regions, with each cell having a particular preferred orientation and spatial frequency^{40–42} (Fig. 2c). These features are also clearly apparent in the model RFs (Fig. 2d).

Unlike previous models^{7,32,51}, the temporal prediction model captures the temporal asymmetry of real RFs. The RF power is highest near the present and decays into the past (Fig. 2d), as observed in real neurons⁵⁰ (Fig. 2c). Furthermore, simple cell RFs have two types of spatiotemporal structure: space-time separable RFs (Fig. 2cI), whose optimal stimulus resembles a flashing or slowly ramping grating, and space-time inseparable RFs, whose optimal stimulus is a drifting grating⁴¹ (Fig. 2cII). Our model captures this diversity (Fig. 2dI-III separable, Fig. 2dIV-VI inseparable).

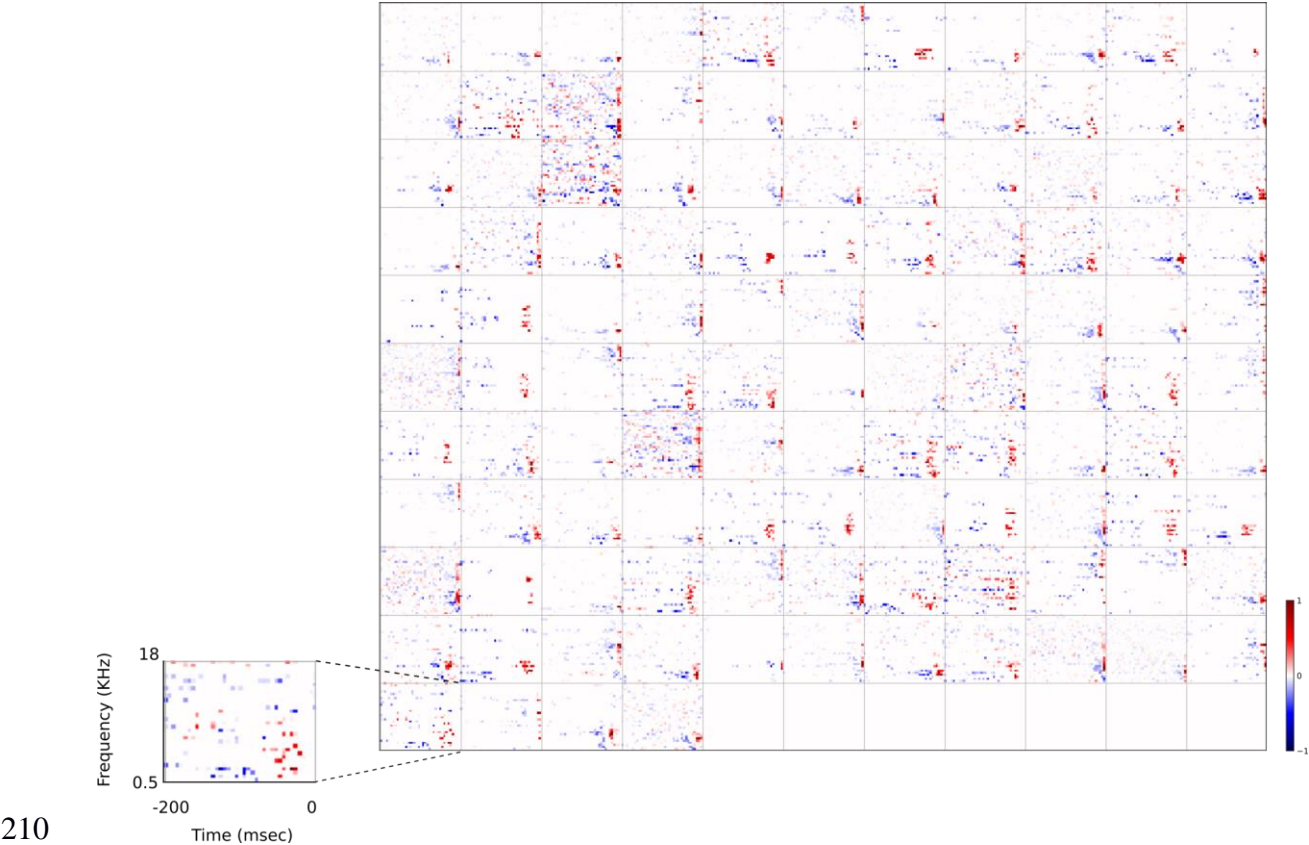
We also examined linear aspects of the tuning of the output units for the visual temporal prediction model using a response-weighted average to white noise input, and found punctate non-oriented RFs that decay into the past.

Qualitative comparison to other models

For comparison, we trained a sparse coding model^{5,6,11} (<https://github.com/zayd/sparsenet>) using our dataset. We would expect such a model to perform less well in the temporal domain, because unlike the temporal prediction model, the direction of time is not explicitly accounted for. The sparse coding model was chosen because it has set the standard for normative models of visual RFs^{5–7,43}, and the same model has also been applied for auditory RFs^{11,49,52,53}. Past studies^{5,6,11} have largely analysed the basis functions produced by the sparse coding model and compared their properties to neuronal RFs. To be consistent with these studies we have done the same, and to have

205 a common term, refer to the basis functions as RFs (although strictly, they are projective fields). We
206 can visually compare the large set of RFs recorded from A1 neurons (Fig. 3) to the full set of RFs
207 obtained from the temporal prediction model when trained on auditory inputs (Fig. 4) and those of
208 the sparse coding model (Fig. 5) when trained on the same auditory inputs.

209



210

211

212 **Figure 3 | Full dataset of real auditory RFs.** 114 neuronal RFs in A1 and AAF of 5 ferrets. Red –
213 excitation, blue - inhibition. Inset shows axes.

214

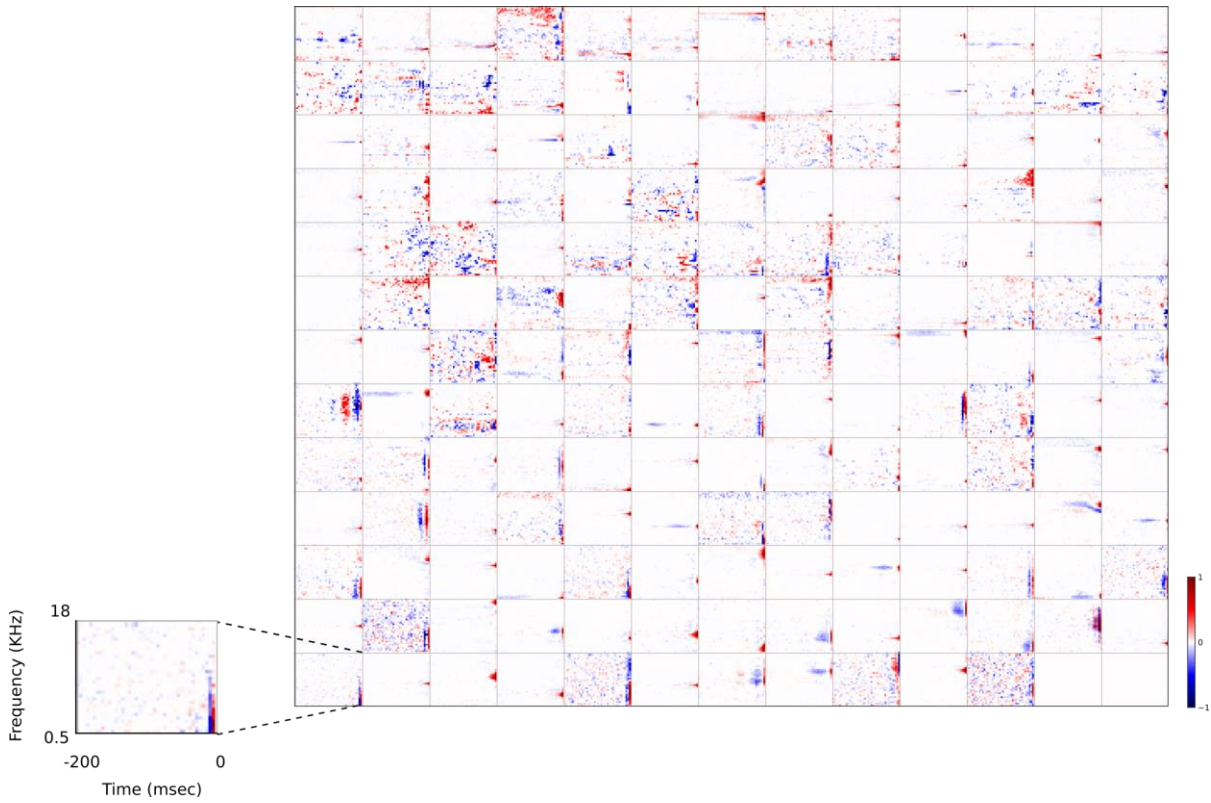


Figure 4 | Full set of auditory RFs of the temporal prediction model units. Units were obtained by training the model with 1600 hidden units on auditory inputs. The hidden unit number and L1 weight regularization strength ($10^{-3.5}$) was chosen because it results in the lowest MSE on the prediction task, as measured using a cross validation set. Many hidden units' weight matrices decayed to near zero during training (due to the L1 regularization), leaving 167 active units. Inactive units were excluded from analysis and are not shown. Example units in Figure 2 come from this set. Red – excitation, blue - inhibition. Inset shows axes. Figure Supplement 1 shows the same RFs on a finer timescale. The full sets of visual spatial and corresponding spatiotemporal RFs for the temporal prediction model when it is trained on visual inputs are shown in Figure Supplements 2-3. Figure Supplement 4 shows the auditory RFs of the temporal prediction model when a linear activation function instead of a sigmoid nonlinearity was used. Figure Supplements 5-7 show the auditory spectrotemporal and visual spatial and 2D spatiotemporal RFs of the temporal prediction model when it was trained on inputs without added noise.

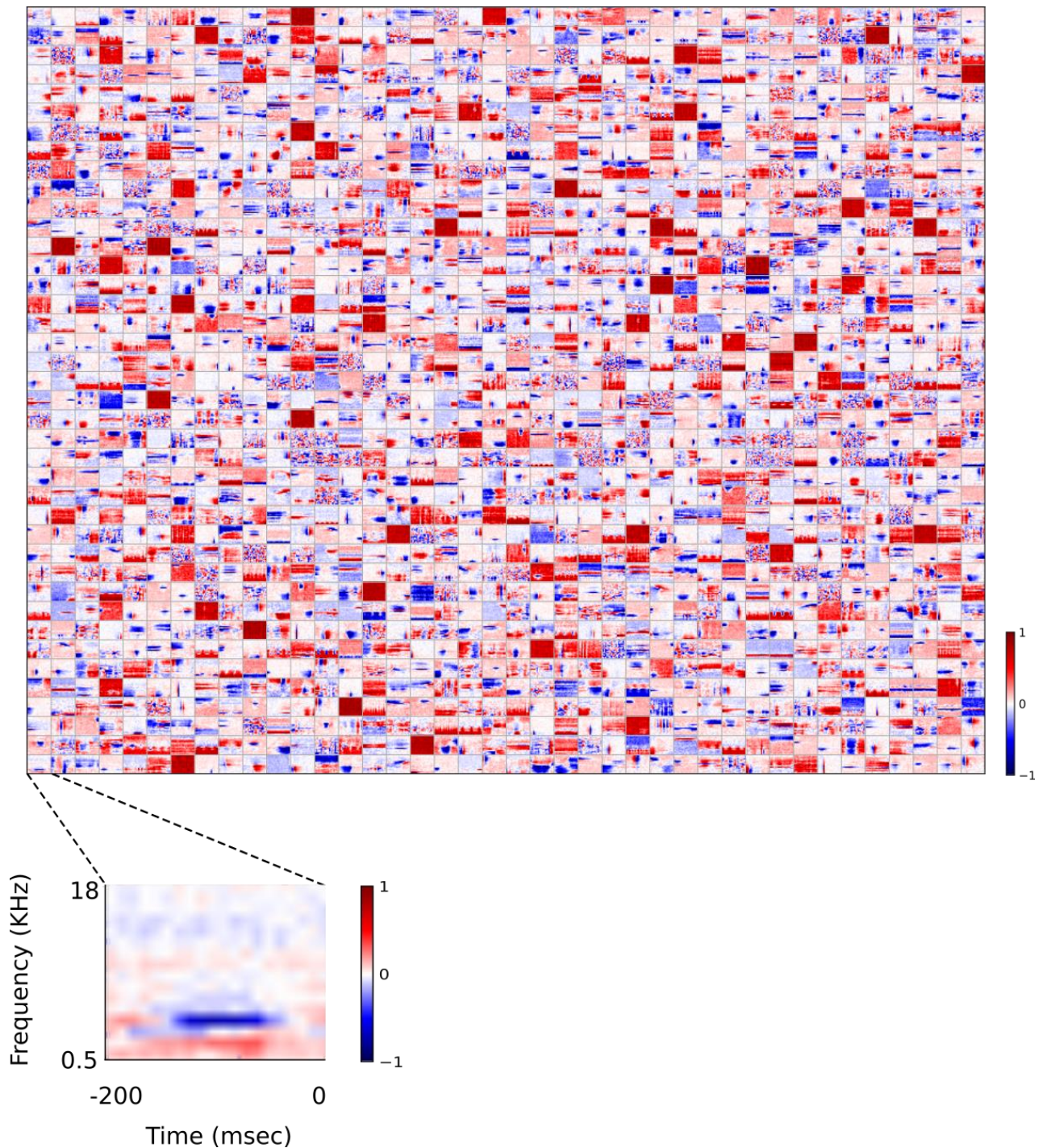


Figure 5 | Full set of auditory ‘RFs’ (basis functions) of sparse coding model used as a control. Units were obtained by training the sparse coding model with 1600 units on the identical auditory inputs used to train the network shown in Figure 4. L1 regularization of $10^{0.5}$ was applied to the units’ activities. This network configuration was selected as it produced unit RFs that most closely resembled those recorded in A1, as determined by visual inspection. Although the basis functions of the sparse coding model are not receptive fields, but projective fields, they tend to be similar in structure^{5,6}. In this manuscript, to have a common term between models and the data, we refer to sparse coding basis functions as RFs. Red – excitation, blue - inhibition. Inset shows axes. The full sets of visual spatial and corresponding spatiotemporal RFs for the sparse coding model when it is trained on visual inputs are shown in Figure Supplements 1-2. Figure Supplements 3-5 show the auditory spectrotemporal and visual spatial and 2D spatiotemporal RFs of the sparse coding model when it was trained on inputs without added noise

A range of RFs were produced by the sparse coding model, some of which show characteristics reminiscent of A1 RFs, particularly in the frequency domain. However, the temporal properties of A1 neurons are not well captured by these RFs. While some RFs display excitation followed by lagging inhibition, very few, if any, show distinct brief excitation followed by extended inhibition. Instead, neurons that show both excitation and inhibition tend to have a symmetric envelope and these features are randomly localised in time, and many neurons display temporally elongated structures that are not found in A1 neurons.

We also trained the sparse coding model on the dataset of visual inputs to serve as a control for the temporal prediction model trained on these same inputs. We compared the full population of spatial and 2D spatiotemporal visual RFs of the temporal prediction model (Fig. 4-Fig. Supplements 2-3) and the sparse coding model (Fig. 5-Fig. Supplements 1-2). As shown in previous studies^{5-7,43}, the sparse coding model produces RFs whose spatial structure resembles that of V1 simple cells (Fig. 5-Fig. Supplements 1-2), but does not capture the asymmetric nature of the temporal tuning of V1 neurons. Furthermore, while it does produce examples of both separable and inseparable spatiotemporal RFs, those that are separable tend to be completely stationary over time, resembling immobile rather than flashing gratings (Fig. 5-Fig. Supplement 2).

Quantitative analysis of auditory results

We compared the RFs generated by both models to the RFs of the population of real A1 neurons we recorded. We first compared the RFs in a non-parametric manner by measuring the Euclidean distances between the coefficient values of the RFs, and then using multi-dimensional scaling to embed these distances in a two-dimensional space (Fig. 6a). The RFs of the sparse coding model span a much larger region than the real A1 and temporal prediction model RFs. Furthermore, the A1 and temporal prediction model RFs occupy a similar region of the space, indicating their greater similarity to each other relative to those of the sparse coding model. We then examined specific

272 attributes of the RFs to determine points of similarity and difference between each of the models
273 and the recorded data. We first considered the temporal properties of the RFs and found that for the
274 data and the temporal prediction model, most of the power is contained in the most recent time-
275 steps (Figs. 2a-b, 3-4, 6b, and Fig. 4-Fig. Supplement 1). Given that the direction of time is not
276 explicitly accounted for in the sparse coding model, as expected, it does not show this feature (Figs.
277 5, 6b). Next, we examined the tuning widths of the RFs in each population for both time and
278 frequency, looking at excitation and inhibition separately. In the time domain, the real data tend to
279 show leading excitation followed by lagging inhibition of longer duration (Figs. 2a, 3, 6c-e). The
280 temporal prediction model also shows many neurons with this temporal structure, with lagging
281 inhibition of longer duration than the leading excitation (Figs. 2b, 4, 6c-e, and Fig. 4-Fig.
282 Supplement 1). This is not the case with the sparse coding model, where units tend to show either
283 excitation and inhibition having the same duration or an elongated temporal structure that does not
284 show such stereotyped polarity changes (Figs. 5, 6c-e). It is also the case that the absolute
285 timescales of excitation and inhibition match more closely in the case of the temporal prediction
286 model (Fig. 6c-e), although a few units display inhibition of a longer duration than is seen in the
287 data (Fig. 6c). The sparse coding model shows a wide range of temporal spans of excitation and
288 inhibition, in keeping with previous studies^{11,48}.

289 Regarding the spectral properties of real neuronal RFs, the spans of inhibition and excitation
290 over sound frequency tend to be similar (Fig. 6f-h). This is also seen in the temporal prediction
291 model, albeit with slightly more variation (Fig. 6f-h). The sparse coding model shows more
292 extensive variation in frequency spans than either the data or our model (Fig. 6f-h).

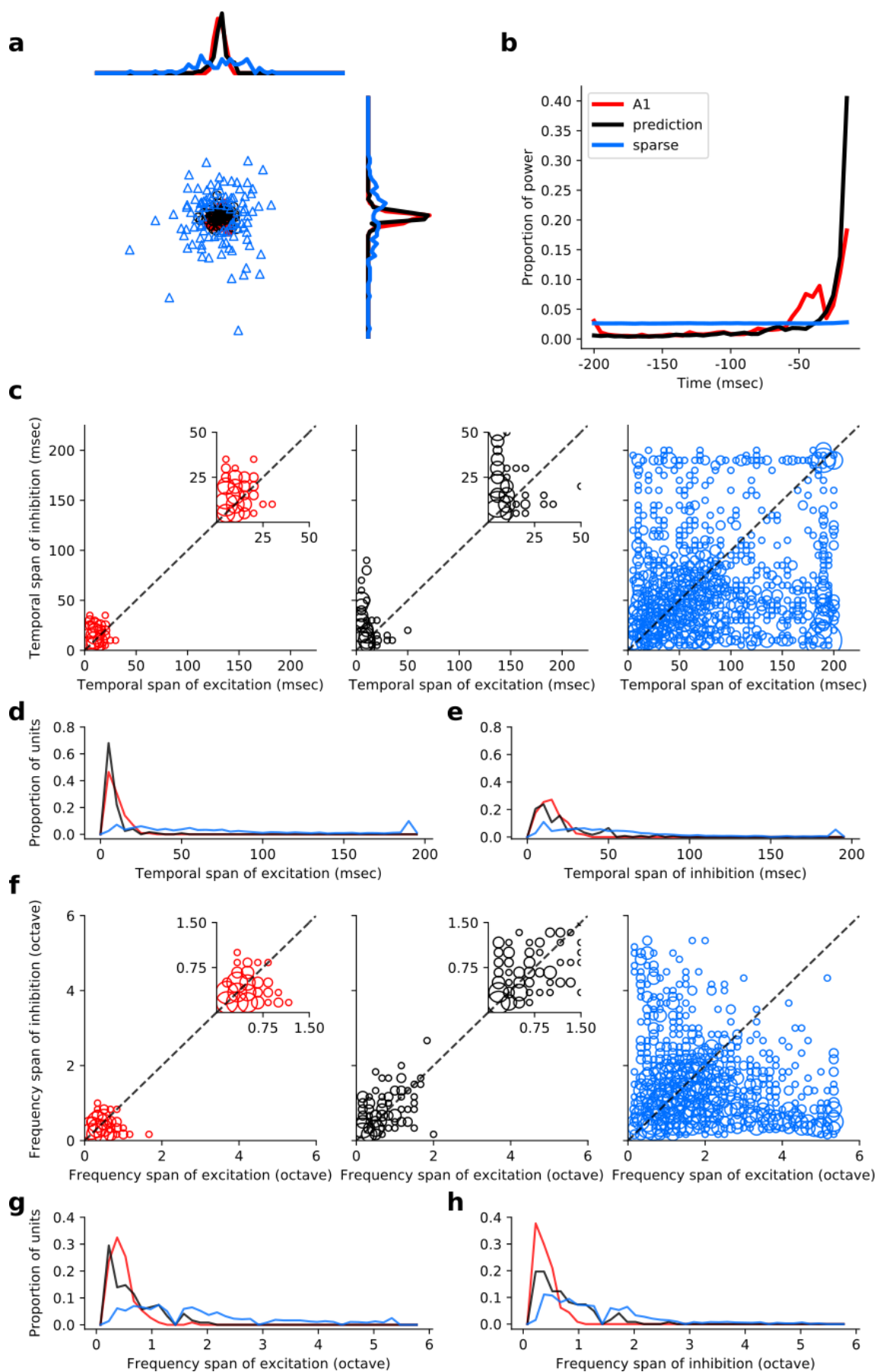


Figure 6 | Population measures for real A1 plus temporal prediction and sparse coding model auditory spectrotemporal RFs. The population measures are taken from the RFs shown

in Figures 3-5. **a**, Each point represents a single RF (with 32 frequency and 38 time steps) which has been embedded in a 2-dimensional space using Multi-Dimensional Scaling (MDS). Red circles - real A1 neurons, black circles – temporal prediction model units, blue triangles – sparse coding model units. Colour scheme applies to all subsequent panels. **b**, Proportion of power contained in each time step of the RF, taken as an average across the population of units. **c**, Temporal span of excitatory subfields versus that of inhibitory subfields, for real neurons and temporal prediction and sparse coding model units. The area of each circle is proportional to the number of occurrences at that point. The inset plots, which zoom in on the distribution use a smaller constant of proportionality for the circles to make the distributions clearer. **d**, Distribution of temporal spans of excitatory subfields, taken by summing along the x-axis in **c**. **e**, Distribution of temporal spans of inhibitory subfields, taken by summing along the y-axis in **c**. **f**, Frequency span of excitatory subfields versus that of inhibitory subfields, for real neurons and temporal prediction and sparse coding model units. **g**, Distribution of frequency spans of excitatory subfields, taken by summing along the x-axis in **f**. **h**, Distribution of frequency spans of inhibitory subfields, taken by summing along the y-axis in **f**. Figure Supplement 1 shows the same analysis for the temporal prediction model and sparse coding model trained on auditory inputs without added noise.

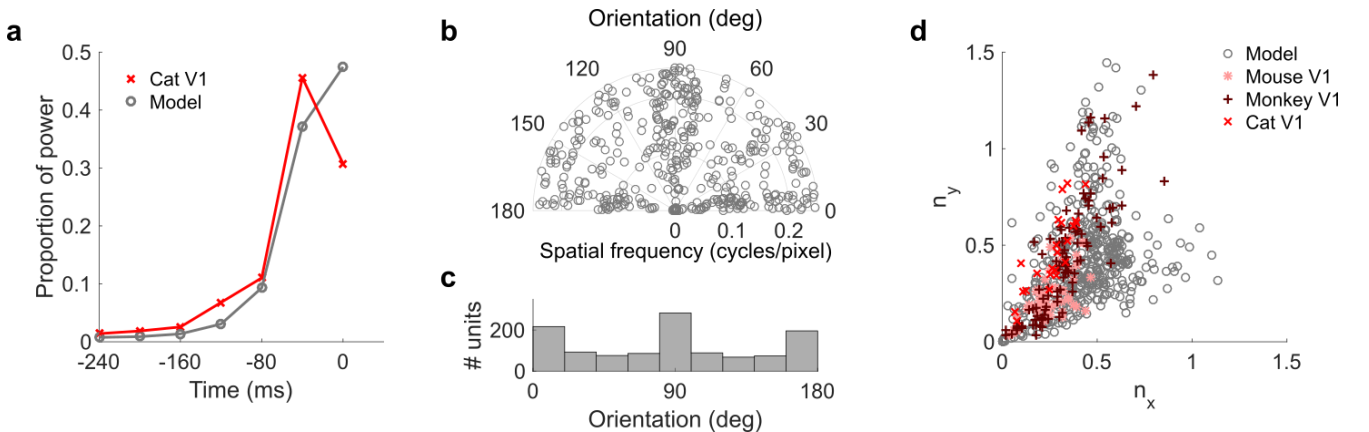
Quantitative analysis of visual results

We also compared the spatiotemporal RFs derived from the temporal prediction and sparse coding models with restricted published datasets summarizing RF characteristics of V1 neurons⁴² and a small number of full spatiotemporal visual RFs acquired from Ohzawa et al⁵⁰. We assessed the orientation and spatial frequency tuning properties of the models' RFs by fitting Gabor functions to their RFs (see Methods).

We compared temporal properties of the RFs from the neural data and the temporal prediction model. In both cases, most power (sum of square of RFs) is in the most recent time steps (Fig. 7a). Previous normative models of spatiotemporal RFs^{7,32,51} (Fig. 7-Fig. Supplement 1c-d) do not show this property, being either invariant over time or localised, but with a symmetric profile that is not restricted to the recent past. We also measured the space-time separability of the RFs of the temporal prediction model (see Methods); substantial numbers of both space-time separable and inseparable units were apparent (631 separable, 969 inseparable; Fig. 4-Fig. Supplement 3). In addition to this, we measured the tilt direction index (TDI) of the model units from their 2D spatiotemporal RFs. This index indicates spatiotemporal asymmetry in space-times RFs and correlates with direction selectivity.^{41,54–57} The mean TDI for the population was 0.33 (0.29 SD),

comparable with the ranges in the neural data (mean 0.16; 0.12 SD in cat area 17/18⁵⁶, mean 0.51; 0.30 SD in macaque V1⁵⁷). Finally, we observed an inverse correlation ($r^2 = -0.31$, $p < 10^{-9}$, $n = 1205$) between temporal and spatial frequency tuning (See Methods), which is also a property of real V1 RFs⁴¹ and is seen in a sparse-coding-related model⁷.

The spatial tuning characteristics of the temporal prediction model's RFs displayed a wide range of orientation and spatial frequency preferences, consistent with the neural data^{41,58} (Fig. 7b-c, Fig. 4-Fig. Supplement 2). Both model and real RFs²⁶ show a preference for spatial orientations along the horizontal and vertical axes, although this orientation bias is seen to a greater extent in the temporal prediction model than in the data. The orientation and frequency tuning characteristics are also well captured by sparse coding related models of spatiotemporal RFs^{7,51} (Fig. 7-Fig. Supplement 1e). Furthermore, the widths and lengths of the RFs of the temporal prediction model, relative to the period of their oscillation, also match the neural data well (Fig. 7d). The distribution of units extends along a curve from blob-like RFs, which lie close to the origin in this plot, to stretched RFs with several subfields, which lie further from the origin. A small proportion of the population have RFs with several short subfields, forming a wing from the main curve in Fig. 7d. Although this property is again fairly well captured by previous models^{5,6,9,42,43} (Fig. 7-Fig. Supplement 1f), only the temporal prediction model seems to be able to capture the blob-like RFs that form a sizeable proportion of the neural data⁴² (Fig. 7d where n_x and $n_y < \sim 0.25$, Fig. 4-Fig. Supplement 2).



351

352 **Figure 7 | Population measures for real V1 and temporal prediction model visual spatial and**
353 **spatiotemporal RFs.** Model units were obtained by training the model with 1600 hidden units on
354 visual inputs. The hidden unit number and L1 weight regularization strength ($10^{-3.75}$) was chosen
355 because it results in the lowest MSE on the prediction task, as measured using a cross validation
356 set. Example units in Figure 2 come from this set. **a**, Proportion of power (sum of squared weights
357 over space and averaged across units) in each time step, for real⁵⁰ and model populations. **b**, Joint
358 distribution of spatial frequency and orientation tuning for population of model unit RFs at their time
359 step with greatest power. **c**, Distribution of orientation tuning for population of model unit RFs at
360 their time step with greatest power. **d**, Distribution of RF shapes for real neurons (cat⁴⁰, mouse⁵⁹
361 and monkey⁴²) and model units. n_x and n_y measure RF span parallel and orthogonal to orientation
362 tuning, as a proportion of spatial oscillation period⁴². For **b-d**, only units that could be well
363 approximated by Gabor functions ($n = 1205$ units; see Methods) were included in the analysis. Of
364 these, only model units that were space-time separable ($n = 473$) are shown in **d** to be comparable
365 with the neuronal data⁴². A further 4 units with $1.5 < n_y < 3.1$ are not shown in **d**. Figure
366 Supplements 1-3 show example visual RFs and the same population measures for the sparse
367 coding model trained on visual inputs with added noise and for the temporal prediction and sparse
368 coding models trained on visual inputs without added noise.

369

370 **Optimising predictive capacity**

371 Under our hypothesis of temporal prediction, we would expect that the better the temporal
372 prediction model network is at predicting the future, the more the RFs of the network should
373 resemble those of real neurons. To examine this hypothesis, we plotted the prediction error of the
374 network as a function of two hyperparameters; the regularisation strength and the number of hidden
375 units (Fig. 8a). Then, we plotted the similarity between the auditory RFs of real A1 neurons and
376 those of the temporal prediction model (Fig. 8b), as measured by the mean KS distances of the
377 temporal and frequency span distributions (Fig. 6d-e, g-h, Methods). The set of hyperparameter
378 settings that give good predictions are also those where the temporal prediction model produces RFs
379 that are most similar to those recorded in A1 ($r^2 = 0.8$, $p < 10^{-9}$, $n = 55$). This result argues that
380 cortical neurons are indeed optimised for temporal prediction.

381

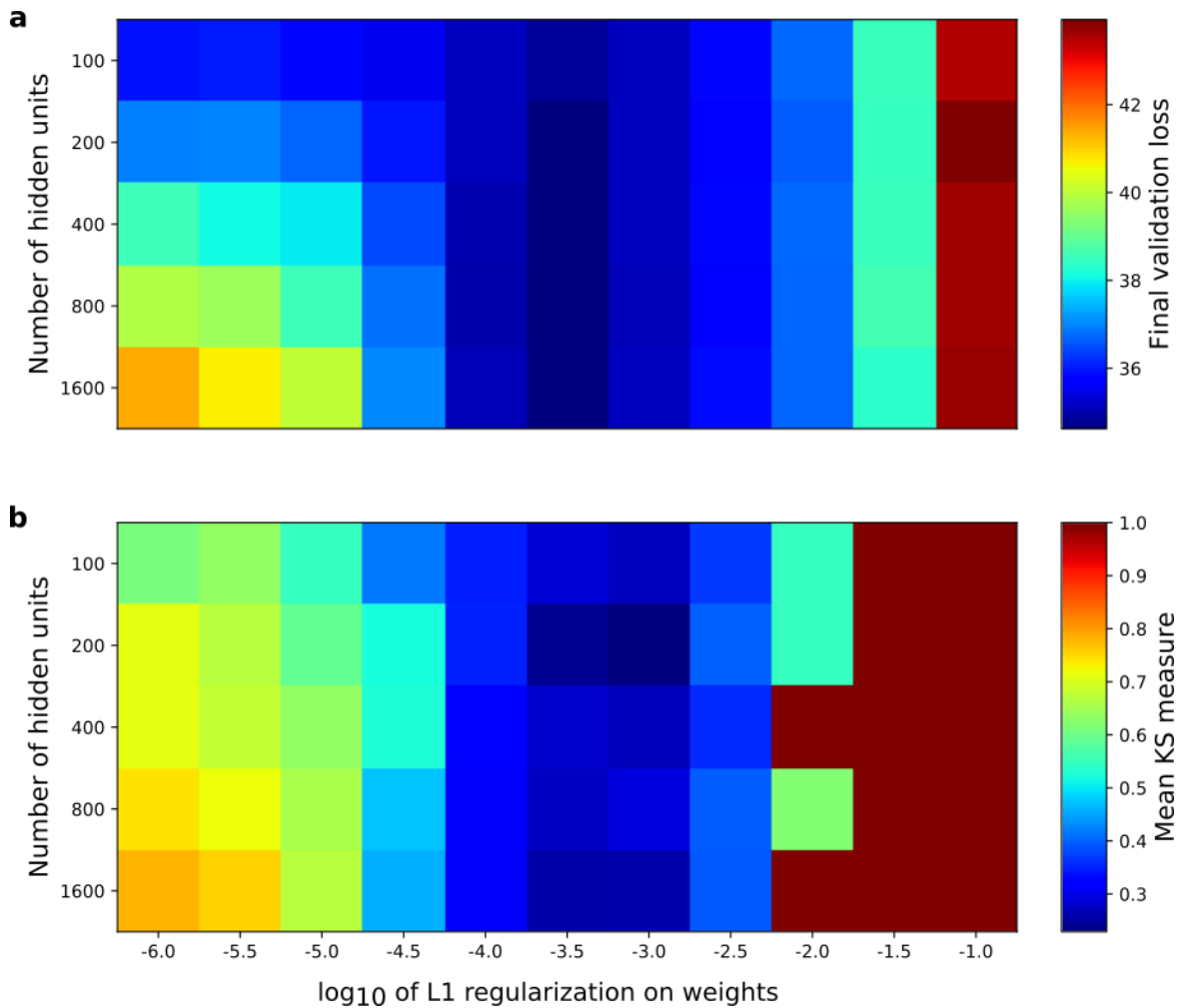


Figure 8 | Correspondence between the temporal prediction model’s ability to predict future auditory input and the similarity of its units’ responses to those of real A1 neurons. Performance of model as a function of number of hidden units and regularization on the weights as measured by **a**, prediction error on validation set at the end of training and **b**, similarity between model units and real A1 neurons. The similarity between the real and model units is measured by averaging the Kolmogorov-Smirnov distance between each of the real and model distributions for the span of temporal and frequency tuning of the excitatory and inhibitory RF subfields (e.g. the distributions in Fig. 6d-e and Fig. 6g-h). Figure Supplement 1 shows the same analysis, performed for the sparse coding model, which does not produce a similar correspondence.

When the similarity measure was examined as a function of the same hyperparameters for the sparse coding model (Fig. 8-Fig. Supplement 1), and this was compared to that model’s stimulus reconstruction capacity as a function of the same hyperparameters, a monotonic relationship between stimulus reconstruction capacity and similarity of real RFs was not found (Fig. 8-Fig. Supplement 1; $r^2 = -0.05$, $p = 0.69$, $n = 50$). In previous studies in which comparisons have been made between normative models and real data, the model hyperparameters have been selected

to maximise the similarity between the real and model RFs. In contrast, the temporal prediction model provides an independent criterion, the prediction error, to perform hyperparameter selection. To our knowledge, no such effective, measurable, independent criterion for hyperparameter selection has been proposed for other normative models of RFs.

Variants of the temporal prediction model

The change in the qualitative structure of the RFs as a function of the number of hidden units and L1 regularization strength, for both the visual and auditory models, can be seen in the interactive supplementary figures (Figure 8-Figure supplements 2-3; https://yossing.github.io/temporal_prediction_model/figures/interactive_supplementary_figures.htm)

1) The main effect of the regularization is to restrict the RFs in space for the visual case and in frequency and time for the auditory case. When the regularization is non-existent or substantially weaker than the optimum for prediction, the visual RFs become less localized in space with more elongated bars. The auditory RFs become more disordered, losing clear structure in most cases. When the regularization is made stronger than the optimum, the RFs become more punctate, for both the visual and auditory models. When the regularization strength is at the optimum for prediction, the auditory and visual model RFs qualitatively most closely resemble those of A1 neurons and V1 simple cells, respectively. This is consistent with what we found quantitatively in the previous section for the auditory model.

The temporal prediction model and the sparse coding model both produce oriented Gabor-like RFs when trained on visual inputs. This raises the possibility that optimization for prediction implicitly optimizes for a sparse response distribution, and hence leads to oriented RFs. To test for this, we measured the sparsity of the temporal prediction model's hidden unit activities (by the Vinje-Gallant measure⁵⁶) in response to the natural image validation set. Examining the relationship between predictive capacity and sparsity, over the range of L1 weight regularization strength and hidden units explored, we did not find a clear monotonic relationship. Indeed, in both the auditory

427 and visual cases, the hidden unit and L1 regularization combination with the best prediction had
428 intermediate sparsity. For the visual case, the best-predicting model had sparsity 0.25, and other
429 models within the grid search had sparsity ranging from 0.16 to 0.57. For the auditory case, the
430 best-predicting model had sparsity 0.58, and other models had sparsity ranging from 0.42 to 0.69.

431 We also varied other characteristics of the temporal prediction model to understand their
432 influence. For both the auditory and visual models, when a different hidden unit nonlinearity (tanh
433 or rectified linear) was used, the networks had similar predictive capacity and produced comparable
434 RFs. However, when the temporal prediction model had linear hidden units, it no longer predicted
435 as well and produced RFs that were less like real neurons in their structure. For the auditory model,
436 the linear model RFs generally became more narrowband in frequency with temporally extended
437 excitation, instead of extended lagging inhibition (Figure 4 - Figure Supplement 4). For the visual
438 model, the linear model RFs showed substantially less similarity to the V1 data. At low
439 regularisation (the best predicting case), the RFs formed full-field grid-like structures. At higher
440 regularisation, they were more punctate, with some units having oriented RFs with short subfields.
441 The RFs also did not change form or polarity over time, but simply decayed into the past.

442 The temporal prediction model and sparse coding model results shown in the main figures of
443 this paper were trained on inputs with added Gaussian noise (6dB SNR), mimicking inherent noise
444 in the nervous system. To determine the effect of adding this noise, all models were also trained
445 without noise, producing similar results (Figure 4 - Figure Supplements 5-7; Figure 5 - Figure
446 Supplements 3-5; Figure 6 - Figure Supplement 1; Figure 7 - Figure Supplements 2-3). The results
447 were also robust to changes in the duration of the temporal window being predicted. We trained the
448 auditory model to predict a span of either 1, 3, 6, or 9 time steps into the future and the visual model
449 to predict 1, 3 or 6 time steps into the future. For the auditory case, we found that increasing the
450 number of time steps being predicted had little effect on the RF structure, both qualitatively and by
451 the KS measure of similarity to the real data. In the visual case, Gabor-like units were present in all
452 cases. Increasing the number of time steps made the RFs more restricted in space and increased the

453 proportion of blob-like RFs.

454

455 **Discussion**

456 We hypothesized that finding features that can efficiently predict future input from its past is a
457 principle that influences the structure of sensory RFs. We implemented an artificial neural network
458 model that instantiates a restricted version of this hypothesis. When this model was trained using
459 natural sounds, it produced RFs that are both qualitatively and quantitatively similar to those of A1
460 neurons. Similarly, when we trained the model using natural movies it produced RFs with many of
461 the properties of V1 simple cells. This similarity is particularly notable in the temporal domain; the
462 model RFs have asymmetric envelopes, with a preference for the very recent past, as is seen in A1
463 and V1. Finally, the more accurate a temporal prediction model is at prediction, the more its RFs
464 tend to be like real neuronal RFs by the measures we use for comparison.

465

466 **Relationship to other models**

467 A number of principles, often acting together, have been proposed to explain the form and diversity
468 of sensory RFs. These include efficient coding^{4-6,11,12,27,49,51,60}, sparseness^{5-7,11,13,43,49,51}, and
469 slowness^{32,48}. Efficient coding indicates that sensory input should be represented as accurately as
470 possible given certain constraints, such as spike count or energy costs. Sparseness posits that only a
471 small proportion of neurons in the population should be active for a given input. Finally, slowness
472 means that neurons should be sensitive to features that change slowly over time. The temporal
473 prediction principle we describe here provides another unsupervised objective of sensory coding. It
474 has been described in a very general manner by the information bottleneck concept^{2,21,28}. We have
475 instantiated a specific version of this idea, with linear-nonlinear encoding of the input, followed by
476 a linear transform from the encoding units' output to the prediction.

477 In the following discussion, we describe previous normative models that infer RFs with

temporal structure from auditory or movie input and relate them to spectrotemporal RFs in A1 or simple cell spatiotemporal RFs in V1, respectively. For focus, other normative models of less directly relevant areas, such as spatial receptive fields without a temporal component^{5,6}, complex cells⁸, retinal receptive fields^{22,27}, or auditory nerve impulse responses⁶¹, will not be examined.

Auditory normative models

A number of coding objectives have been explored in normative models of A1 spectrotemporal RFs. One approach¹² found analytically that the optimal typical spectrotemporal RF for efficient coding was spectrally localised with lagging and flanking inhibition, and showed an asymmetric temporal envelope. However, the resulting RF also showed substantially more flanking inhibition, more ringing over time and frequency, and operated over a much shorter timescale (~10ms) than seen in A1 RFs (Fig. 3). Moreover, this approach produced a single generic RF, rather than capturing the diversity of the population.

Other models have produced a diverse range of spectrotemporal RFs. In the sparse coding approach^{11,49,52,53}, a spectrogram snippet is reconstructed from a sum of basis functions (a linear generative model), each weighted by its unit's activity, with a constraint to have few active units. This approach is the same as the sparse coding model we used as a control (Fig. 5). A challenge with many sparse generative models is that the activity of the units is found by a recurrent iterative process that needs to find a steady state; this is fine for static stimuli such as images, but for dynamic stimuli like sounds it is questionable whether the nervous system would have sufficient time to settle on appropriate activities before the stimulus had changed. Related work also used a sparsity objective, but rather than minimising stimulus reconstruction error, forced high dispersal¹³ or decorrelation^{10,48} of neural responses. Although lacking some of the useful probabilistic interpretations of sparse generative models, this approach does not require a settling process for inference. An alternative to sparseness is temporal slowness, which can be measured by temporal coherence⁴⁸. Here the linear transform from sequential spectrogram snippets to unit activity is

504 optimised to maximise the correlation of each unit's response over a certain time window, while
505 maintaining decorrelation between the units' activities.

506 Although the frequency tuning derived with these models can resemble that found in the
507 midbrain or cortex^{10,11,13,48,49,52,53} (Fig. 5), the resulting RFs lack the distinct asymmetric temporal
508 profile and lagging inhibition seen in real midbrain or A1 RFs. Furthermore, they often have
509 envelopes that are too elongated over time, often spanning the full temporal width of the
510 spectrotemporal RF. This is related to the fact that the time window to be encoded by the model is
511 set arbitrarily, and every time point within that window is given equal weighting, i.e. the direction
512 of time is not accounted for. This is in contrast to the temporal prediction model, which naturally
513 gives greater weighting to time-points near the present than the past due to their greater predictive
514 capacity.

515

516 **Visual normative models**

517 The earliest normative model of spatiotemporal RFs of simple cells used independent component
518 analysis (ICA)⁷, which is practically equivalent for visual or auditory data to the critically complete
519 case of the sparse coding model^{5,6} we used as a control (Fig. 5-Fig. Supplements 1-2 and Fig. 7-
520 Fig. Supplement 1). The RFs produced by this model and our control model reproduced fairly well
521 the spatial aspects of simple cell RFs. However, in contrast to the temporal prediction model, the
522 subset of more 'blob-like' RFs seen in the data are not well captured by our control sparse coding
523 model (Fig. 7d and Fig. 7-Fig. Supplement 1f). In the temporal domain, again unlike the temporal
524 prediction model and real V1 simple cells, the RFs of the ICA and sparse coding model are not
525 pressed up against the present with an asymmetrical temporal envelope, but instead show a
526 symmetrical envelope or span the entire range of times examined. A related model⁵¹ assumes that a
527 longer sequence of frames is generated by convolving each basis function with a time-varying
528 sparse coefficient and summing the result, so that each basis function is applied at each point in
529 time. The resulting spatiotemporal RFs are similar to those produced by ICA⁷, or our control model

530 (Fig. 5-Fig. Supplement 2 and Fig. 7-Fig. Supplement 1c). Although they tend not to span the entire
531 range of times examined, they do show a symmetrical envelope, and require an iterative inference
532 procedure, as described above for audition.

533 Temporal slowness constraints have also been used to model the spatiotemporal RFs of
534 simple cells. The bubbles³² approach combines sparse and temporal coherence constraints with
535 reconstruction. The resulting RFs show similar spatial and temporal properties to those found using
536 ICA. A related framework is slow feature analysis (SFA)^{8,62}, which enforces temporal smoothness
537 by minimizing the derivative of unit responses over time, while maximising decorrelation between
538 units. SFA has been used to model complex cell spatiotemporal RFs (over only two time steps⁸),
539 and a modified version has been used to model spatial (not spatiotemporal) RFs of simple cells⁹.
540 These results are not directly comparable with our results or the spatiotemporal RFs of simple cells.

541 In the slowness framework, the features found are those that persist over time; the presence
542 of such a feature in the recent past predicts that the same feature will be present in the near future.
543 This is also the case for our predictive approach, which, additionally, can capture features in the
544 past that predict features in the future that are subtly or radically different from themselves. The
545 temporal prediction principle will also give different weighting to other features, as it values
546 predictive capacity rather than temporal slowness³⁰. In addition, although slowness models can be
547 extended to model RFs over more than one time step^{8,32,48}, capturing temporal structure, they do not
548 inherently give more weighting to information in the most recent past and therefore do not give rise
549 to asymmetric temporal profiles in RFs.

550 There is one study that has directly examined temporal prediction as an objective for visual
551 RFs in a manner similar to ours⁶³. Here, as in our model, a single hidden layer feedforward neural
552 network was used to predict the immediate future frame of a movie patch from its past frames.
553 However, only two frames of the past were used in this study, so a detailed exploration of the
554 temporal profile of the spatiotemporal RFs was not possible. Nevertheless, some similarities and
555 differences in the spatial RFs between the two frames were noted, and some units had oriented RFs.

556 In contrast to our model, however, many RFs were noisy and did not resemble those of simple cells.
557 Potential reasons for this difference include the use of L_2 rather than L_1 regularization on the
558 weights, an output nonlinearity not present in our model, the optimization algorithm used, network
559 size, or the dataset. Another very recent related study²⁰ also implemented a somewhat different form
560 of temporal prediction, with a linear (rather than linear-nonlinear) encoder, and linear decoder.
561 When applied to visual data oriented receptive fields were produced, but they were spatio-
562 temporally separable and hence not direction selective.

563 564 **Strengths and limitations of the temporal prediction model**

565 Temporal prediction has several strengths as an objective function for sensory processing. First, it
566 can capture underlying features in the world²; this is also the case with sparseness^{5,6} and slowness⁶²,
567 but temporal prediction will prioritise different features. Second, it can predict future inputs, which
568 is very important for guiding action, especially given internal processing delays. Third, objectives
569 such as efficient or sparse reconstruction retain everything about the stimulus, whereas an important
570 part of neural information processing is the selective elimination of irrelevant information⁶⁴.
571 Prediction provides a good initial criterion for eliminating potentially unwanted information.
572 Fourth, prediction provides a natural method to determine the hyperparameters of the model (such
573 as regularization strength, number of hidden units, activation function and temporal window size).
574 Other models select their hyperparameters depending on what best reproduces the neural data,
575 whereas we have an independent criterion – the capacity of the network to predict the future. One
576 notable hyperparameter is how many time-steps of past input to encode. As described above, this is
577 naturally decided by our model because only time-steps that help predict the future have significant
578 weighting. Fifth, the temporal prediction model computes neuronal activity without needing to
579 settle to a steady state, unlike some other models^{5–7,11,43,49,52,53}. For dynamic stimuli, a model that
580 requires settling may not reach equilibrium in time to be useful. Sixth, and most importantly,
581 temporal prediction successfully models many aspects of the RFs of primary cortical neurons. In

582 addition to accounting for spatial and spectral tuning in V1 and A1, respectively, at least as well as
583 other normative models, it reproduces the temporal properties of RFs, particularly the asymmetry of
584 the envelopes of RFs, something few previous models have attempted to explain.

585 Although the temporal prediction model's ability to describe neuronal RFs is high, the
586 match with real neurons is not perfect. For example, the span of frequency tuning of our modelled
587 auditory RFs is narrower than in A1 (Fig 6g-h). We also found an overrepresentation of vertical and
588 horizontal orientations compared to real V1 data (Fig 7b-c). Some of these differences could be a
589 consequence of the data used to train the model. Although the statistics of natural stimuli are
590 broadly conserved⁶⁵, there is still variation⁶⁶, and the dataset used to train the network may not
591 match the sensory world of the animal experienced during development and over the course of
592 evolution. In future work, it would be valuable to explore the influence of natural datasets with
593 different statistics, and also to match those datasets more precisely to the evolutionary context and
594 individual experience of the animals examined. Furthermore, a comparison of the model with neural
595 data from different species, at different ages, and reared in different environments would be useful.

596 Another cause of differences between the model and neural RFs may be the recording
597 location of the RFs and how they are characterised. We used the primary sensory cortices as regions
598 for comparison, because we performed transformations on the input data that are similar to the
599 preprocessing that takes place in afferent subcortical structures. We spatially filtered the visual data
600 in a similar way to the retina^{5,6}, and spectrally decomposed the auditory data as in the inner ear, and
601 then used time bins (5ms) which are coarser than, but close to, the maximum amplitude modulation
602 period that can be tracked by auditory midbrain neurons⁶⁷. However, primary cortex is not a
603 homogenous structure, with neurons in different layers displaying certain differences in their
604 response properties⁶⁸. Furthermore, the methods by which neurons are sampled from the cortex may
605 not provide a representative sample. For example, multi-electrode arrays tend to favour larger and
606 more active neurons. In addition, the method and stimuli used to construct RFs from the data can
607 bias their structure somewhat⁴⁵.

608 The model presented here is based on a simple feedforward network with one layer of
609 hidden units. This limits its ability to predict features of the future input, and to account for RFs
610 with nonlinear tuning. More complex networks, with additional layers or recurrency may allow the
611 model to account for more complex tuning properties, including those found beyond the primary
612 sensory cortices. Careful, principled adjustment of the preprocessing, or different regularisation
613 methods (such as sparseness or slowness applied to the units' activities), may also help. There is an
614 open question as to whether the current model may eliminate some information that is useful for
615 reconstruction of the past input or for prediction of higher order statistical properties of the future
616 input, which might bring it into conflict with the principle of least commitment⁶⁹. It is an empirical
617 question how much organisms preserve information that is not predictive of the future, although
618 there are theoretical arguments against such preservation². Such conflict might be remedied, and the
619 model improved, by adding feedback from higher areas or by adding an objective^{4-6,60} to
620 reconstruct the past or present in addition to predicting the future.

621 To determine whether the model could help explain neuronal responses in higher areas, it
622 would be useful to develop a hierarchical version of the temporal prediction model, applying the
623 same model again to the activity of the hidden units rather than to the input. Another useful
624 extension would be to see if the features learnt by the temporal prediction model could be used to
625 accelerate learning of useful tasks such as speech or object recognition, by providing input or
626 initialisation for a supervised or reinforcement learning network. Indeed, temporal predictive
627 principles have been shown to be useful for unsupervised training of networks used in visual object
628 recognition⁷⁰⁻⁷³.

629 Finally, it is interesting to consider possible more explicit biological bases for our model.
630 We envisage the input units of the model as thalamic input, and the hidden units as primary cortical
631 neurons. Although the function of the output units could be seen as just a method to optimize the
632 hidden units to find the most predictive code given sensory input statistics, they may also have a
633 physiological analogue. Current evidence⁷⁴⁻⁷⁶ suggests that while primary cortical RFs are to an

634 extent hard-wired in form by natural selection, their tuning is also refined by individual sensory
635 experience. This refinement process may require a predictive learning mechanism in the animal's
636 brain, at least at some stage of development and perhaps also into adulthood. Hence, one might
637 expect to find a subpopulation of neurons that represent the prediction (analogous to the output
638 units of the model) or the prediction error (analogous to the difference between the output unit
639 activity and the target). Indeed, signals relating to sensory prediction error have been found in
640 A1⁷⁷, though they may also be located in other regions of the brain. Finally, it is important to note
641 that, although the biological plausibility of backpropagation has long been questioned, recent
642 progress has been made in developing trainable networks that perform similarly to artificial neural
643 networks trained with backpropagation, but with more biologically plausible characteristics⁷⁸, for
644 example, by having spikes or avoiding the weight transport problem⁷⁹.

645

646 **Conclusion**

647 We have shown that a simple principle - predicting the imminent future of a sensory scene from its
648 recent past - explains many features of the RFs of neurons in both primary visual and auditory
649 cortex. This principle may also account for neural tuning in other sensory systems, and may prove
650 useful for the study of higher sensory processing and aspects of neural development and learning.
651 While the importance of temporal prediction is increasingly widely recognized, it is perhaps
652 surprising nonetheless that many basic tuning properties of sensory neurons, which we have known
653 about for decades, appear, in fact, to be a direct consequence of the brain's need to efficiently
654 predict what will happen next.

655

656 **Methods**

657 **Data used for model training and testing**

658 *Visual inputs*

Videos (without sound, sampled at 25 fps) of wildlife in natural settings were used to create visual stimuli for training the artificial neural network. The videos were obtained from <http://www.arkive.org/species>, contributed by: BBC Natural History Unit, <http://www.gettyimages.co.uk/footage/bbcmotiongallery>; BBC Natural History Unit & Discovery Communications Inc., <http://www.bbcmotiongallery.com>; Granada Wild, <http://www.itnsource.com>; Mark Deeble & Victoria Stone Flat Dog Productions Ltd., <http://www.deeblestone.com>; Getty Images, <http://www.gettyimages.com>; National Geographic Digital Motion, <http://www.ngdigitalmotion.com>. The longest dimension of each video frame was clipped to form a square image. Each frame was then band-pass filtered⁶ and downsampled (using bilinear interpolation) over space, to provide 180x180 pixel frames. Non-overlapping patches of 20x20 pixels were selected from a fixed region in the centre of the frames, where there tended to be visual motion. The video patches were cut into sequential overlapping clips each of 8 frames duration. Thus, each training example (clip) was made up of a 20x20 pixel section of the video with a duration of 8 frames (320ms), providing a training set of $N = \sim 500,000$ clips from around 5.5 h of video, and a validation set of $N = \sim 100,000$ clips. Finally, the training and validation sets were normalised by subtracting the mean and dividing by the standard deviation (over all pixels, frames and clips in the training set). The goal of the neural network was to predict the final frame (the ‘future’) of each clip from the first 7 frames (the ‘past’).

677

Auditory inputs

Auditory stimuli were compiled from databases of human speech (~60%), animal vocalizations (~20%) and sounds from inanimate objects found in natural settings (e.g. running water, rustling leaves; ~20%). Stimuli were recorded using a Zoom H4 or collected from online sources. Natural sounds were obtained from www.freesound.org, contributed by users sedi, higginsdj, jult, kvgarlic, xenognosis, zabuhailo, funnyman374, videog, j-zazvurek, samueljustice00, gfrog, ikbenraar, felix-blume, orbitalchiller, saint-sinner, carlvus, kyser, vflefevre, hitrison, willstepp, timbahrij,

685 xdimebagx, r-nd0mm3m, the-yura, rsilveira-88, stomachache, foongaz, edufigg, yurkobb,
686 sander motions, darius-kedros, freesoundjon-01, dwightsabeast, borralbi, acclivity, J.Zazvurek,
687 Zabuhailo, soundmary, Darius Kedros, Kyster, urupin, RSilveira and freelibras. Human speech
688 sounds were obtained from <http://databases.forensic-voice-comparison.net/>^{80,81}.

689 Each sound was sampled at (or resampled to) 44.1 kHz and converted into a simple
690 ‘cochleagram’, to make it more analogous to the activity pattern that would be passed to the
691 auditory pathway after processing by the cochlea. To calculate the cochleagram, a power
692 spectrogram was computed using 10ms Hanning windows, overlapping by 5ms (giving time steps
693 of 5ms). The power across neighbouring Fourier frequency components was then aggregated into
694 32 frequency channels using triangular windows with a base width of 1/3 octave whose centre
695 frequencies ranged from 500 to 17,827Hz (1/6th octave spacing, using code adapted from
696 melbank.m, <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>). The cochleagrams were
697 then decomposed into sequential overlapping clips, each of 43 time steps (415 ms) in duration,
698 providing a training set of ~1,000,000 clips (~1.3 hours of audio) and a validation set of ~200,000
699 clips. To approximately model the intensity compression seen in the auditory nerve⁸², each
700 frequency band in the stimulus set was divided by the median value in that frequency band over the
701 training set, and passed through a hill function, defined as $h(x) = cx/(1+cx)$ with $c=0.02$. Finally, the
702 training and cross-validation sets were normalised by subtracting the mean and dividing by the
703 standard deviation over all time steps, frequency bands and clips in the training set. The first 40
704 time steps (200 ms) of each clip (the ‘past’) were used as inputs to the neural network, whose aim
705 was to predict the content (the ‘future’) of the remaining 3 time steps (15 ms).

706

707 *Addition of Gaussian noise*

708 To replicate the effect of noise found in the nervous system, Gaussian noise was added to both the
709 auditory and visual inputs with a signal-to-noise ratio (SNR) of 6dB. While the addition of noise did
710 not make substantial differences to the RFs of units trained on visual inputs, this improved the

711 similarity to the data when the model was trained on auditory inputs. The results from training the
 712 network on inputs without added noise are shown for auditory inputs in Fig. 4-Fig. Supplement 5
 713 and Fig. 6-Fig. Supplement 1 and for visual inputs in Fig. 4-Fig. Supplement 6-7 and Fig. 7-Fig.
 714 Supplement 2. The results from the sparse coding model were similar in both cases for inputs with
 715 and without noise (Figs. 5-6, Fig. 5-Fig. Supplements 1-5, Fig. 6-Fig. Supplement 1, Fig. 7-Fig.
 716 Supplements 1,3).

717

718 **Temporal prediction model**

719 *The model and cost function*

720 The temporal prediction model was implemented using a standard fully connected feed-forward
 721 neural network with one hidden layer. Each hidden unit in the network computed the sum of
 722 linearly weighted inputs, and its output was determined by passing this sum through a monotonic
 723 nonlinearity. This nonlinearity $s = f(a)$ was either a logistic function $f(a) = 1/(1+\exp(-a))$ or a
 724 similar nonlinear function (such as \tanh). For results reported here, we used the logistic function,
 725 though obtained similar results when we trained the model using $f(a) = \tanh(a)$. For comparison, we
 726 also trained the model replacing the nonlinearity with a linear function, where $f(a) = a$. In this case,
 727 we found that the RFs tended to be punctate in space or frequency and did not typically show the
 728 alternating excitation and inhibition characteristic real neurons in A1 and V1.

729 Formally, for a network with $i = 1$ to I input variables, $k = 1$ to K output units and a single
 730 layer of $j = 1$ to J hidden units, the output s_{jn} of hidden unit j for clip n is given by:

731

$$s_{jn} = f\left(b_j + \sum_{i=1}^I w_{ji} u_{in}\right)$$

732 (1)

733 The value u_{in} of input variable i for clip n is simply the value for a particular pixel and time
 734 step (frame) of the ‘past’ in preprocessed visual clip n ($I = 20 \text{ pixels} \times 20 \text{ pixels} \times 7 \text{ time steps} = 2800$),

735 or the value for a particular frequency band and time step of the ‘past’ of cochleagram clip n ($I = 32$
736 frequencies \times 40 time steps= 1280). Hence, the index i spans over several frequencies or pixels and
737 also over time steps into the past. The subscript n has been dropped for clarity in the figures (Fig.
738 1). The parameters in Equation 1 are the connective input weights w_{ji} (between each input variable i
739 and hidden unit j), and the bias b_j (of hidden unit j).

740 The activity \hat{v}_{kn} of each output unit k , which is the estimate of the true future v_{kn} given the
741 past u_{in} , is given by:

742

$$\hat{v}_{kn} = b_k + \sum_{j=1}^J w_{kj} s_{jn}$$

743 (2)

744 The parameters in Equation 2 are the connective output weights w_{kj} (between each hidden
745 unit j and output unit k) and the bias b_k (of output unit k). The activity \hat{v}_{kn} of output unit k for clip n
746 is the estimate for a particular pixel of the ‘future’ in the visual case ($K = 20 \times 20 \times 1 = 400$), or the
747 value for a particular frequency band and time step of the ‘future’ in the auditory case ($K = 32 \times 3 =$
748 96).

749 The parameters w_{ji} , w_{kj} , b_j , and b_k were optimised for the training set by minimizing the cost
750 function given by:

751

$$E = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K (\hat{v}_{kn} - v_{kn})^2 + \lambda \left(\sum_{i=1}^I \sum_{j=1}^J |w_{ji}| + \sum_{j=1}^J \sum_{k=1}^K |w_{kj}| \right)$$

752 (3)

753 Thus, E is the sum of the squared error between the prediction \hat{v}_{kn} and the target v_{kn} over all N
754 training examples, plus an L_1 regularisation term, which is proportional to the sum of absolute
755 values of all weights in the network and its strength is determined by the hyper-parameter λ . This
756 regularisation tends to drive redundant weights to near zero and provides a parsimonious network.

757

758 *Implementation details*

759 The networks were implemented in Python ([https://lasagne.readthedocs.io/en/latest/](https://lasagne.readthedocs.io/en/latest/http://deeplearning.net/software/theano/);
760 <http://deeplearning.net/software/theano/>). The objective function was minimised using
761 backpropagation as performed by the Adam optimisation method⁸³. An alternative implementation
762 of the model was also made in MATLAB using the Sum-of-Functions Optimiser⁸⁴
763 (<https://github.com/Sohl-Dickstein/Sum-of-Functions-Optimizer>) to train the network using
764 backpropagation. Training examples were split into minibatches of approximately 7000 training
765 examples each.

766 During model network training, several hyperparameters were varied, including the
767 regularisation strength (λ), the number of units in the hidden layer and the nonlinearity used by each
768 hidden unit. For each hyperparameter setting, the training algorithm was run for 1000 iterations.
769 Running the network for longer (10000 iterations) showed negligible improvement to the prediction
770 error (as measured on the validation set) or change in RF structure.

771 The effect of varying the number of hidden units and λ on the prediction error for the
772 validation set is shown in Fig. 8. In both the visual and auditory case, the results presented are the
773 networks that predicted best on the validation set after 1000 iterations through the training data. For
774 the auditory case, the settings that resulted in the best prediction were 1600 hidden units and $\lambda =$
775 $10^{-3.5}$, while in the visual case, the optimal settings were 1600 hidden units and $\lambda = 10^{-6.25}$.

776

777 *Model receptive fields*

778 In the model, the combination of linear weights and nonlinear activation function are similar to the
779 basic linear non-linear (LN) model^{85–89} commonly used to describe neural RFs. Hence, the input
780 weights between the input layer and a hidden unit of the model network are taken directly to
781 represent the unit's RF, indicating the features of the input that are important to that unit.

782 Because of the symmetric nature of the sigmoid function, $f(a) = 1-f(-a)$, with appropriate

783 modification of the biases, a hidden unit has the same influence on the prediction if its input and
784 output matrices are both multiplied by -1. That is, for unit j , if we convert w_{ij} to $-w_{ij}$, w_{jk} to $-w_{jk}$, b_j to
785 $-b_j$, and b_k to $-b_k + w_{jk}$, this will have no effect on the prediction or the cost function. This can be
786 done independently for each hidden unit. Hence, the sign of each unit's RF could equally be
787 positive or negative and have the same result on the predictions given by the network. However, we
788 know that auditory units always have leading excitation (Fig. 3). Hence, for both the predictive
789 model and for the sparse coding model, we assume leading excitation for each unit. This was done
790 for all analyses.

791 As more units are added to the model network, the number of inactive units increases. To
792 account for this, we measured the relative strength of all input connections to each hidden unit by
793 summing the square of all input weights for that unit. Units for which the sum of square input
794 weights was $<1\%$ of the maximum strength for the population were deemed to be inactive and
795 excluded from all subsequent analyses. The difference in connection strength between active and
796 inactive units was very distinct; a threshold $<0.0001\%$ only marginally increases the number of
797 active units.

798

799 **Sparse coding model**

800 The sparse coding model was used as a control for both visual and auditory cases. The Python
801 implementation of this model (<https://github.com/zayd/sparsenet>) was trained using the same visual
802 and auditory inputs used to train the predictive model. The training data were divided into mini-
803 batches which were shuffled and the model optimised for one full pass through the data. Inference
804 was performed using the Fast Iterative Shrinkage and Thresholding (FISTA) algorithm. A sparse L_1
805 prior with strength λ , was applied to the unit activities. A range of λ -values and unit numbers were
806 tried (Fig. 8-Fig. Supplement 1). The learning rate and batch size were also varied until reasonable
807 values were found. As there was no independent criterion by which to determine the 'best' settings,
808 we chose the network that produced basis functions whose receptive fields were most similar to

those of real neurons. In the auditory case, this was determined using the mean KS measure of similarity (Figure 8-Figure Supplement 1). In the visual case, as a similarity measure was not performed, this was done by inspection. In both cases, the model configurations chosen were restricted to those trained in an overcomplete condition (having more units than the number of input variables) in order to remain consistent with previous instantiations of this model^{5,6,11}. In this manner, we selected a sparse coding network with 1600 units, $\lambda=10^{0.5}$, learning rate = 0.01 and 100 mini-batches in the auditory case (Figs. 5-6). In the visual case, the network selected was trained with 3200 units, $\lambda=10^{0.5}$, learning rate = 0.05 and 100 mini-batches (Fig. 5-Fig. Supplements 1-2 and Fig. 7-Fig. Supplement 1). Although these sparse coding basis functions are projective fields, they tend to be similar in structure to receptive fields^{5,6}, and, for simplicity, are referred to as RFs.

Auditory receptive field analysis

In vivo A1 RF data

Auditory RFs of neurons were recorded in the primary auditory cortex (A1) and anterior auditory field (AAF) of 5 pigmented ferrets of both sexes (all > 6 months of age) and used as a basis for comparison with the RFs of model units trained on auditory stimuli. Systematic differences in response properties of A1 and AAF neurons are minor and not relevant for this study, and for simplicity here, we refer to neurons from either primary field indiscriminately as “A1 neurons”. These recordings were performed under license from the UK Home Office and were approved by the University of Oxford Committee on Animal Care and Ethical Review. Full details of the recording methods are described in earlier studies^{45,90}. Briefly, we induced general anaesthesia with a single intramuscular dose of medetomidine ($0.022 \text{ mg} \cdot \text{kg}^{-1} \cdot \text{h}^{-1}$) and ketamine ($5 \text{ mg} \cdot \text{kg}^{-1} \cdot \text{h}^{-1}$), which was then maintained with a continuous intravenous infusion of medetomidine and ketamine in saline. Oxygen was supplemented with a ventilator, and we monitored vital signs (body temperature, end-tidal CO_2 , and the electrocardiogram) throughout the experiment. The temporal muscles were retracted, a head holder was secured to the skull surface,

835 and a craniotomy and a durotomy were made over the auditory cortex. Extracellular recordings
836 were made using silicon probe electrodes (Neuronexus Technologies) and acoustic stimuli were
837 presented via Panasonic RPHV27 earphones, which were coupled to otoscope specula that were
838 inserted into each ear canal, and driven by Tucker-Davis Technologies System III hardware (48
839 kHz sample rate).

840 The neuronal recordings used the “BigNat” stimulus set⁴⁵, which consists of natural sounds
841 including animal vocalizations (e.g., ferrets and birds), environmental sounds (e.g., water and
842 wind), and speech. To identify those neural units that were driven by the stimuli, we calculated a
843 “noise ratio” statistic^{89,91} for each units and excluded from further analysis any neurons with a noise
844 ratio >40. In total, driven spiking responses of 114 units (75 single unit, 39 multi-unit) were
845 recorded to this stimulus set. Then, the auditory (spectrotemporal) RF of each unit was constructed
846 using a previously described method⁴⁵. Briefly, linear regression was performed in order to
847 minimise the squared error between each neuron’s spiking response over time and the cochleagram
848 of the stimuli that gave rise to that response. The method used was exactly the same as in our earlier
849 study⁴⁵, except that L_1 rather than L_2 regularisation was used to constrain the regression. The
850 spectrotemporal RFs of these neurons took the same form as the inputs to the model neural network
851 (i.e., 32 frequencies and 40 time-steps over the same range of values) and were therefore
852 comparable to the model units’ RFs. In order to account for the latency of auditory cortical
853 responses, the first two time-steps (10ms) of the neuronal responses were removed, leaving 38 time-
854 steps.

855

856 *Multi-Dimensional Scaling (MDS)*

857 To get a non-parametric indication of how similar the model units’ RFs were to those of real A1
858 neurons, each RF was embedded into a 2-dimensional space using MDS (Fig. 6a and Fig. 6-Fig.
859 Supplement 1a). First, 100 units each from the temporal prediction and sparse coding models and
860 from the real population were chosen at random. To ensure that the model RFs were of the same

861 dimensionality prior to embedding, the final two time steps of each model RF were removed.

862

863 *Measuring temporal and frequency spans of RFs*

864 We quantified the span, over time and frequency, of the excitatory and inhibitory subfields of each
865 RF. To do this, each RF was first separated into excitatory and inhibitory subfields, where the
866 excitatory subfield was the RF with negative values set to 0, and the inhibitory subfield the RF with
867 positive values set to 0. In some cases, model units did not exhibit notable inhibitory subfields. To
868 account for this, the power contained in each subfield was calculated (sum of the squares of the
869 subfield). Inhibitory subfields with $< 5\%$ of the power of that unit's excitatory subfield were
870 excluded from further analysis. According to this criterion, 44 of 167 active units in the temporal
871 prediction model and 193 of 1600 units in the sparse model did not display inhibition.

872 Singular value decomposition (SVD) was performed on each subfield separately, and the
873 first pair of singular vectors was taken, one of which is over time, the other over frequency. For the
874 excitatory subfield, the temporal span was measured as the proportion of values in the temporal
875 singular vector that exceeded 50% of the maximum value in the vector. The same analysis provided
876 the temporal span for the inhibitory subfield. Similarly, we measured the frequency spans of the
877 RFs by applying this measure to the frequency singular vectors of the excitatory and the inhibitory
878 subfields.

879 We also examined, for both real and model RFs, the mean power for each of the 38 time
880 steps in the RFs (Fig. 6b), which was calculated as the mean of the squared RF values, over all
881 frequencies and RFs, at each time step.

882

883 *Mean KS measure*

884 To compare each network's units with those recorded in A1 (Fig. 3), the two-sample Kolmogorov-
885 Smirnov (KS) distance between the real and model distribution was measured for both the temporal
886 and spectral span of the excitatory and inhibitory subfields (e.g. the distributions in Fig. 6d-e and

Fig. 6g-h). These four KS measures were then averaged to give a single mean KS measure for each network, indicating how closely the temporal and frequency characteristics of real and model units matched on average for that network. The KS measure is low for similar distributions and high for distributions that diverge greatly. Thus networks whose units display temporal and frequency tuning characteristics that match those of real neurons more closely give rise to a lower mean KS measure.

Visual receptive field analysis

In vivo V1 RF data

Visual RFs measured using recordings from V1 simple cells were compared against the model (Fig. 2c, and Fig. 7a (cat⁵⁰) and Fig. 7d (cat⁴⁰, mouse⁵⁹ and monkey⁴²)). The in vivo data were taken from the authors' website¹⁶ or extracted from relevant papers⁴⁰ or provided by the authors^{50,59}.

Fitting Gabors

In order to quantify tuning properties of the model's visual RFs, 2D Gabors were fitted to the optimal time-step of each unit's response^{40,42}. This allowed comparison to previous experimental studies which parameterised real RFs by the same method⁴². The optimal time-step was defined⁴² as the time-step of the unit's response which contained the most power (sum of square values). The Gabor function has been shown to provide a good approximation for most spatial aspects of simple visual RFs^{40,42}. The 2D Gabor is given as:

$$G(x', y') = A \exp \left(- \left(\frac{x'}{\sqrt{2}\sigma_x} \right)^2 - \left(\frac{y'}{\sqrt{2}\sigma_y} \right)^2 \right) \cos(2\pi f x' + \phi) \quad (6)$$

where, the spatial coordinates (x', y') are acquired by translating the centre of the RF (x_0, y_0) to the origin and rotating the RF by its spatial orientation θ :

$$x' = (x - x_0) \cos\theta + (y - y_0) \sin\theta \quad (7)$$

$$y' = -(x - x_0) \sin\theta + (y - y_0) \cos\theta \quad (8)$$

σ_x and σ_y provide the width of the Gaussian envelope in the x' and y' directions, while f and ϕ parameterise the spatial frequency and phase of the sinusoid along the x' axis. A parameterises the height of the Gaussian envelope.

For each RF, the parameters ($x_0, y_0, \sigma_x, \sigma_y, \theta, f, \phi$) of the Gabor were fitted by minimizing the mean squared error between the Gabor model and the RF using the minFunc minimization package (<http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html>). In order to avoid local minima, the fitting was performed in two steps. First, the spatial RF was converted to the spectral domain using a 2D Fourier transform. Since the Fourier transform of a 2D Gabor is a 2D Gaussian⁴⁰, which is easier to fit, an estimate of many of the parameters was obtained by first fitting a 2D Gaussian in the spectral domain. Using the parameters obtained from the spectral fitting as initial estimates, a 2D Gabor was then fitted to the original RF in the spatial domain. The fitted parameters provided a good estimate of the units' responses, with residual errors between the spatial responses and the corresponding Gabor fits being small and lacking spatial structure, and the median pixel-wise correlation coefficient of the Gabor fits for the temporal prediction model units was 0.88. Units whose fitted Gabors had a poor fit (those with a correlation coefficient <0.7 ; 214 units) were excluded from further analysis. We also excluded units with a high correlation coefficient (>0.7) if the centre position of the Gabor was estimated to be outside the RF, and hence only the Gabor's tail was being fitted to the response (39 units), and those for which the estimated standard deviation of the Gaussian envelope in either x or y was <0.5 pixels, which meant very few non-negligible pixel values were used to constrain the parameters (146 units). Together, these exclusion criteria (which sometimes overlapped), led to 395 of the 1600 responsive units being excluded for the temporal prediction model.

937

938 *2D spatiotemporal receptive fields*

939 In order to better view their temporal characteristics we collapsed the 3D spatiotemporal real and
940 model RFs (space-space-time) along a single spatial direction to create 2D spatiotemporal (space-
941 time) representations⁴¹. First, we determined the 3D RFs' optimal time step (the time step with the
942 largest sum of squared values). We then acquired the rotation and translation that centres the RF on
943 zero and places the oriented bars parallel to the y-axis at the optimal time step from the Gabor
944 parameterisation of each unit at its optimal time step. We applied this fixed transformation to each
945 time step and collapsed the RF by summing the activity along the newly defined y-axis. The
946 resulting 2D (space-time) RFs provide intuitive visualisation of the RF across time, while losing
947 minimal information. For the RFs of real neurons⁵⁰, the most recent time step (40ms) of the 3D and
948 2D spatiotemporal RFs were removed to account for the latency of V1 neurons (Fig 2c; 7a).

949

950 *Estimating space-time separability*

951 The population of model units contained both space-time (ST) separable and inseparable units. First
952 the two spatial dimensions of the 20x20x7 3D RF were collapsed to a single vector to yield a single
953 400x7 matrix. The SVD of this matrix was then taken and the singular values examined. If the ratio
954 between the second and first singular value was ≥ 0.5 , the unit was deemed to be inseparable.
955 Otherwise, the unit was deemed to be separable. Examining the 20x7 2D spatiotemporal RFs
956 (obtained as outlined in the preceding section; Fig. 4-Fig. Supplement 3) showed this to be an
957 accurate way of separating space-time separable and inseparable units.

958

959 *Spatial RF structure*

960 For comparison with the real V1 RF and previous theoretical studies, the width and length of our
961 model's RFs were measured relative to their spatial frequency⁴². Here, $n_x = \sigma_x f$ gives a measure of
962 the length of the bars in the RF, while $n_y = \sigma_y f$ gives a measure of the number of oscillations of its

963 sinusoidal component. Thus, in the n_y , n_x plane, blob-like RFs with few cycles lie close to the
964 origin, while stretched RFs with many subfields lie away from the origin. RFs with values high
965 along the n_y axis, have many bars, while those far along the n_x axis have long bars. As in
966 Ringach⁴² only space-time separable units were included in this analysis.

967

968 *Temporal weighting profile of the population*

969 The mean power for each of the 7 time steps of the RFs was examined for both real and model
970 populations (Fig. 7a). The temporal weighting profile was calculated as the mean, over space and
971 the population, of the squared values of the 2D spatiotemporal RFs at each time step.

972

973 *Tilt direction index*

974 The tilt direction index (TDI)^{41,54–57} of an RF is given by $(R_p - R_q)/(R_p + R_q)$, where R_p is the
975 amplitude at the peak of the 2D Fourier transform of the 2D spatiotemporal RF, found at spatial
976 frequency F_{space} and temporal frequency F_{time} . R_q is the amplitude at $(F_{\text{space}}, -F_{\text{time}})$ in the 2D Fourier
977 transform. The mean and standard deviations of TDI for experimental data for the cat⁵⁷ and
978 macaque⁵⁷ were measured from data extracted from figures in the relevant references (Figure 11A
979 and the low-contrast axis of Figure 3A respectively).

980

981 *Peak temporal frequency*

982 The 2D spatiotemporal RFs were also useful for calculating further temporal response properties of
983 the model. The temporal frequency was calculated as the peak temporal frequency of each
984 spatiotemporal RF as measured from its 2D Fourier transform.

985

986 **Code and data availability**

987 All custom code used in this study was implemented in MATLAB and Python. We have uploaded
988 the code to a public Github repository⁹². The raw auditory experimental data is available

989 at <https://osf.io/ayw2p/>. The movies and sounds used for training the models are all publicly
990 available at the websites detailed in the Methods.

991

992 **Acknowledgments**

993 Nicol Harper was supported by a Sir Henry Wellcome Postdoctoral Fellowship (WT082692) and
994 other Wellcome Trust funding (WT076508AIA, WT108369/Z/2015/Z), by the Department of
995 Physiology, Anatomy and Genetics at the University of Oxford, by Action on Hearing Loss (PA07),
996 and by the Biotechnology and Biological Sciences Research Council (BB/H008608/1). Yosef
997 Singer and Yayoi Teramoto were supported by the Clarendon Fund. Yayoi Teramoto was supported
998 by the Wellcome Trust (10525/Z/14/Z). Andrew King and Ben Willmore were supported by the
999 Wellcome Trust (WT076508AIA, WT108369/Z/2015/Z). We thank Bruno Olshausen for
1000 discussions on his model.

1001

1002 **References**

- 1003 1. Nijhawan, R. Motion extrapolation in catching. *Nature* **370**, 256–257 (1994).
- 1004 2. Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural*
1005 *Comput.* **13**, 2409–63 (2001).
- 1006 3. Attneave, F. Some informational aspects of visual perception. *Psychol. Rev.* **61**, 183–93
1007 (1954).
- 1008 4. Barlow, H. B. Sensory mechanisms, the reduction of redundancy, and intelligence. in *The*
1009 *Mechanisation of Thought Processes* (eds. Blake, D. V. & Uttley, A. M.) 535–539 (H. M.
1010 Stationary Office, 1959).
- 1011 5. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by
1012 learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
- 1013 6. Olshausen, B. A. & Field, D. J. Sparse coding with an overcomplete basis set: a strategy
1014 employed by V1? *Vision Research* **37**, 3311–3325 (1997).

- 1015 7. van Hateren, J. H. & Ruderman, D. L. Independent component analysis of natural image
1016 sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc.*
1017 *R. Soc. Lond. B* **265**, 2315–2320 (1998).
- 1018 8. Berkes, P. & Wiskott, L. Slow feature analysis yields a rich repertoire of complex cell
1019 properties. *J. Vis.* **5**, 579–602 (2005).
- 1020 9. Berkes, P., Turner, R. E. & Sahani, M. A structured model of video reproduces primary
1021 visual cortical organisation. *PLoS Comput. Biol.* **5**, (2009).
- 1022 10. Klein, D. J., König, P. & Körding, K. P. Sparse spectrotemporal coding of sounds. *EURASIP*
1023 *J. Appl. Signal Processing* **2003**, 1–9 (2003).
- 1024 11. Carlson, N. L., Ming, V. L. & DeWeese, M. R. Sparse codes for speech predict
1025 spectrotemporal receptive fields in the inferior colliculus. *PLoS Comput. Biol.* **8**, (2012).
- 1026 12. Zhao, L. & Zhaoping, L. Understanding auditory spectro-temporal receptive fields and their
1027 changes with input statistics by efficient coding principles. *PLoS Comput. Biol.* **7**, (2011).
- 1028 13. Kozlov, A. S. & Gentner, T. Q. Central auditory neurons have composite receptive fields.
1029 *Proc. Natl. Acad. Sci.* **113**, 1441–1446 (2016).
- 1030 14. Cusack, R. & Carlyon, R. Auditory perceptual organization inside and outside the laboratory.
1031 in *Echological psychoacoustics* (ed. Neuhoﬀ, J. G.) 15–48 (Elsevier, 2004).
- 1032 15. Bixler, E. O., Bartlett, N. R. & Lansing, R. W. Latency of the blink reflex and stimulus
1033 intensity *J. Percept. Psychophys.* **2**, 559–560 (1967).
- 1034 16. Yeomans, J. S. & Frankland, P. W. The acoustic startle reflex: neurons and connections.
1035 *Brain Res. Rev.* **21**, 301–314 (1996).
- 1036 17. Bizley, J. K., Nodal, F. R., Nelken, I. & King, A. J. Functional organization of ferret auditory
1037 cortex. *Cereb. Cortex* **15**, 1637–53 (2005).
- 1038 18. Helmholtz, H. *Concerning the perceptions in general Treatise on Physiological Optics (3rd*
1039 *ed)*.
- 1040 19. Sutton, R. S.; Barton, A. G. An adaptive network that constructs and uses an internal model

of its world. *Cogn. Brain Theory* **4**, 217–246 (1981).

20. Chalk, M., Marre, O. & Tkačik, G. Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 186–191 (2018).

21. Salisbury, J. M. & Palmer, S. E. Optimal prediction in the retina and natural motion statistics. *J. Stat. Phys.* **162**, 1309–1323 (2016).

22. Huang, Y. & Rao, R. P. N. Predictive coding. *Wiley Interdiscip. Rev. Cogn. Sci.* **2**, 580–593 (2011).

23. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).

24. Friston, K. Learning and inference in the brain. *Neural Networks* **16**, 1325–1352 (2003).

25. Rao, R. P. & Ballard, D. H. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Comput.* **9**, 721–63 (1997).

26. Rao, R. P. An optimal estimation approach to visual perception and learning. *Vis. Res* **39**, 1963–89. (1999).

27. Srinivasan, M. V, Laughlin, S. B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. London. Ser. B, Biol. Sci.* **216**, 427–59 (1982).

28. Palmer, S. E., Marre, O., Berry, M. J. & Bialek, W. Predictive information in a sensory population. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6908–13 (2015).

29. Heeger, D. J. Theory of cortical function. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1773–1782 (2017).

30. Creutzig, F. & Sprekeler, H. Predictive coding and the slowness principle: an information-theoretic approach. *Neural Comput.* **20**, 1026–1041 (2008).

31. Kayser, C. *et al.* Extracting Slow Subspaces from Natural Videos Leads to Complex Cells. *Artif. NEURAL NETWORKS – ICANN 2001, Vol. 2130 Lect. NOTES Comput. Sci.* 1075–1080 (2001).

32. Hyvärinen, A., Hurri, J. & Väyrynen, J. Bubbles: a unifying framework for low-level

- statistical properties of natural image sequences. *J. Opt. Soc. Am. A* **20**, 1237–1252 (2003).
33. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2**, 284 (1985).
34. Aertsen, A. M., Olders, J. H. & Johannesma, P. I. Spectro-temporal receptive fields of auditory neurons in the grassfrog. III. Analysis of the stimulus-event relation for natural stimuli. *Biol. Cybern.* **39**, 195–209 (1981).
35. Aertsen, A. M. H. J. & Johannesma, P. I. M. A comparison of the Spectro-Temporal sensitivity of auditory neurons to tonal and natural stimuli. *Biol. Cybern.* **42**, 145–156 (1981).
36. Reid, R. C., Soodak, R. E. & Shapley, R. M. Linear mechanisms of directional selectivity in simple cells of cat striate cortex. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 8740–4 (1987).
37. deCharms, R. C., Blake, D. T. & Merzenich, M. M. Optimizing sound features for cortical neurons. *Science* **280**, 1439–43 (1998).
38. Harper, N. S. *et al.* Network Receptive Field Modeling Reveals Extensive Integration and Multi-feature Selectivity in Auditory Cortical Neurons. *PLOS Comput. Biol.* **12**, e1005113 (2016).
39. Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148**, 574–91 (1959).
40. Jones, J. P. & Palmer, L. A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58**, 1233–1258 (1987).
41. DeAngelis, G. C., Ohzawa, I. & Freeman, R. D. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *J. Neurophysiol.* **69**, 1091–1117 (1993).
42. Ringach, D. L. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.* **88**, 455–463 (2002).
43. van Hateren, J. H. & van der Schaaf, A. Independent component filters of natural images

- compared with simple cells in primary visual cortex. *Proc. R. Soc. London B* **265**, 359–366 (1998).
44. Eliasmith, C. & Anderson, C. H. *Neural engineering: computation, representation, and dynamics in neurobiological systems*. (MIT Press, 2003).
 45. Willmore, B. D. B., Schoppe, O., King, A. J., Schnupp, J. W. H. & Harper, N. S. Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing. *J. Neurosci.* **36**, 280–9 (2016).
 46. Nodal, F. R. & King, A. J. Hearing and Auditory Function in Ferrets. in *Biology and Diseases of the Ferret* 685–710 (John Wiley & Sons, Inc., 2014). doi:10.1002/9781118782699.ch29
 47. Kavanagh, G. L. & Kelly, J. B. Hearing in the ferret (*Mustela putorius*): effects of primary auditory cortical lesions on thresholds for pure tone detection. *J. Neurophysiol.* **60**, 879–888 (1988).
 48. Carlin, M. A. & Elhilali, M. Sustained firing of model central auditory neurons yields a discriminative spectro-temporal representation for natural sounds. *PLoS Comput. Biol.* **9**, (2013).
 49. Brito, C. S. N. & Gerstner, W. Nonlinear Hebbian learning as a unifying principle in receptive field formation. *PLoS Comput. Biol.* **12**, (2016).
 50. Ohzawa, I., DeAngelis, G. C. & Freeman, R. D. Encoding of binocular disparity by simple cells in the cat's visual cortex. *J. Neurophysiol.* **75**, 1779–805 (1996).
 51. Olshausen, B. A. Learning sparse, overcomplete representations of time-varying natural images. *IEEE Int. Conf. Image Process.* (2003).
 52. Młynarski, W. & McDermott, J. H. Learning Mid-Level Auditory Codes from Natural Sound Statistics. (2017).
 53. Blättler, F., Hahnloser, R. H. R., Doupe, A., Hahnloser, R. & Wilson, R. An Efficient Coding Hypothesis Links Sparsity and Selectivity of Neural Responses. *PLoS One* **6**, e25506 (2011).

- 1119 54. Pack, C. C., Conway, B. R., Born, R. T. & Livingstone, M. S. Spatiotemporal structure of
1120 nonlinear subunits in macaque visual cortex. *J. Neurosci.* **26**, 893–907 (2006).
- 1121 55. Anzai, A., Ohzawa, I. & Freeman, R. D. Joint-encoding of motion and depth by visual
1122 cortical neurons: neural basis of the Pulfrich effect. *Nat. Neurosci.* **4**, 513–518 (2001).
- 1123 56. Baker, C. L. Linear filtering and nonlinear interactions in direction-selective visual cortex
1124 neurons: a noise correlation analysis. *Vis. Neurosci.* **18**, 465–85 (2001).
- 1125 57. Livingstone, M. S. & Conway, B. R. Contrast Affects Speed Tuning, Space-Time Slant, and
1126 Receptive-Field Organization of Simple Cells in Macaque V1. *J. Neurophysiol.* **97**, 849–857
1127 (2007).
- 1128 58. Kreile, A. K., Bonhoeffer, T. & Hübener, M. Altered visual experience induces instructive
1129 changes of orientation preference in mouse visual cortex. *J. Neurosci.* **31**, 13911–13920
1130 (2011).
- 1131 59. Niell, C. M. & Stryker, M. P. Highly Selective Receptive Fields in Mouse Visual Cortex. *J.*
1132 *Neurosci.* **28**, (2008).
- 1133 60. Attneave, F. Some informational aspects of visual perception. *Psychol. Rev.* **61**, 183–93
1134 (1954).
- 1135 61. Smith, E. C. & Lewicki, M. S. Efficient auditory coding. *Nature* **439**, 978–982 (2006).
- 1136 62. Wiskott, L. & Sejnowski, T. J. Slow feature analysis: unsupervised learning of invariances.
1137 *Neural Comput.* **14**, 715–770 (2002).
- 1138 63. Palm, R. B. Prediction as a candidate for learning deep hierarchical models of data.
1139 (Technical University of Denmark, (DTU) Informatics, 2012).
- 1140 64. Marzen, S. E. & DeDeo, S. The evolution of lossy compression. *J. R. Soc. Interface* **14**,
1141 (2017).
- 1142 65. Field, D. J. Relations between the statistics of natural images and the response properties of
1143 cortical cells. *J. Opt. Soc. Am. A* **4**, 2379 (1987).
- 1144 66. Torralba, A. & Oliva, A. Statistics of natural image categories. *Comput. Neural Syst* **14**, 391–

412 (2003).

67. Rees, A. & Møller, A. R. Responses of neurons in the inferior colliculus of the rat to AM and FM tones. *Hear. Res.* **10**, 301–330 (1983).

68. Harris, K. D. & Mrsic-Flogel, T. D. Cortical connectivity and sensory coding. *Nature* **503**, 51–58 (2013).

69. Marr, D. Early processing of visual information. *Phil. Trans. R. Soc. Lond. B* **275**, 441–475 (1976).

70. Srivastava, N., Mansimov, E. & Salakhutdinov, R. Unsupervised Learning of Video Representations using LSTMs. *Icml 2009* (2015). doi:citeulike-article-id:13519737

71. Ranzato, M. *et al.* Video (language) modeling: a baseline for generative models of natural videos. (2016).

72. Lotter, W., Kreiman, G. & Cox, D. Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. (2016).

73. Oh, J., Guo, X., Lee, H., Lewis, R. & Singh, S. Action-Conditional Video Prediction using Deep Networks in Atari Games. *Adv. Neural Inf. Process. Syst.* **28**, (2015).

74. Dahmen, J. C. & King, A. J. Learning to hear: plasticity of auditory cortical processing. *Curr. Opin. Neurobiol.* **17**, 456–464 (2007).

75. Huberman, A. D., Feller, M. B. & Chapman, B. Mechanisms underlying development of visual maps and receptive fields. *Annu. Rev. Neurosci.* **31**, 479–509 (2008).

76. Kiorpes, L. Visual development in primates: Neural mechanisms and critical periods. *Dev. Neurobiol.* **75**, 1080–90 (2015).

77. Rubin, J., Ulanovsky, N., Nelken, I. & Tishby, N. The representation of prediction error in auditory cortex. *PLoS Comput. Biol.* **10**, 1–28 (2016).

78. Bengio, Y., Lee, D.-H., Bornschein, J. & Lin, Z. Towards Biologically Plausible Deep Learning. *arXiv Prepr. arxiv1502.0415* 18 (2015). doi:10.1007/s13398-014-0173-7.2

79. Lillicrap, T. P., Cownden, D., Tweed, D. B. & Akerman, C. J. Random synaptic feedback

- weights support error backpropagation for deep learning. *Nat. Commun.* **7**, 13276 (2016).
80. Morrison, G. S. *et al.* Forensic database of voice recordings of 500+ Australian English speakers. (2015).
 81. Morrison, G. S., Rose, P. & Zhang, C. Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Aust. J. Forensic Sci.* **44**, 155–167 (2012).
 82. Sachs, M. B. & Abbas, P. J. Rate versus level functions for auditory-nerve fibers in cats: tone-burst stimuli. *J. Acoust. Soc. Am.* **56**, 1835–47 (1974).
 83. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. (2014).
 84. Sohl-Dickstein, J., Poole, B. & Ganguli, S. Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods. *Proc. 31st Int. Conf. Mach. Learn.* **32**, 604–612 (2014).
 85. Simoncelli, E., Pillow, J. W., Paninski, L. & Schwartz, O. Characterization of neural responses with stochastic stimuli. in *The cognitive neurosciences, III* (ed. Gazzaniga, M.) 327–338 (MIT Press, 2004).
 86. Dahmen, J. C., Hartley, D. E. H. & King, A. J. Stimulus-Timing-Dependent Plasticity of Cortical Frequency Representation. **28**, 13629–13639 (2008).
 87. Craig A. Atencio, Tatyana O. Sharpee, and C. E. S. Nonlinearities in Auditory Cortical Neurons. **58**, 956–966 (2009).
 88. Chichilnisky, E. J. A simple white noise analysis of neuronal light responses. *Network* **12**, 199–213 (2001).
 89. Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H. & King, A. J. Contrast Gain Control in Auditory Cortex. *Neuron* **70**, 1178–1191 (2011).
 90. Bizley, J. K., Walker, K. M. M., Silverman, B. W., King, A. J. & Schnupp, J. W. H. Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *J. Neurosci.* **29**, 2064–75 (2009).

1197 91. Sahani, M. & Linden, J. F. J. F. How linear are auditory cortical responses? *Adv. Neural Inf.*
1198 *Process. Syst.* **15**, 109–116 (2003).

1199 92. Singer, Y. temporal_prediction_model. Github
1200 https://github.com/yossing/temporal_prediction_model. ba8ed26.

1201

1202

1203

Figure supplement legends

Figure 4-Figure Supplement 1 | Full set of auditory RFs of the temporal prediction model units shown on a finer timescale. All details are as in Figure 3, but the only the most recent 100ms of the response profile is shown in order to illustrate details of the RFs. Inset shows axes.

Figure 4-Figure Supplement 2 | Full set of visual spatial RFs of the temporal prediction model units. Model units are the same as those used in Figure 7a. Example units in Fig. 2 come from this set. Each square represents the spatial RF of a single unit, shown at its best time step. The best time step was determined by selecting the time step for which the power (sum of squares) of the RF was greatest. White – excitation, black - inhibition.

Figure 4-Figure Supplement 3 | Visual 2D (space-time) spatiotemporal RFs of temporal prediction model units. Model units are the same as those used in Fig. 7. Each square represents the 2D spatiotemporal RF of a single unit corresponding to the unit in the same position in Fig. 4-Fig. Supplement 2, obtained by summing across space along the axis of the orientation for that unit. Red – excitation, blue - inhibition. Inset shows axes.

Figure 4-Figure Supplement 4 | Full set of auditory RFs of the temporal prediction model units using a linear activation function. Units were obtained by training the model with 1600 hidden units on auditory inputs. The hidden unit number and L1 weight regularization strength ($10^{-3.25}$) was chosen because they result in the lowest MSE on the prediction task, as measured using a cross validation set. Almost all hidden units' weight matrices decayed to near zero during training (due to the L1 regularization), leaving 35 active units. Inactive units were excluded from analysis and are not shown. Red – excitation, blue - inhibition. Inset shows axes.

Figure 4-Figure Supplement 5 | Full set of auditory RFs of the temporal prediction model units trained on auditory inputs without added noise. Units were obtained by training the model with 1600 hidden units on auditory inputs. The hidden unit number and L1 weight regularization strength (10^{-4}) was chosen because it results in the lowest MSE on the prediction task, as measured using a cross validation set. Many hidden units' weight matrices decayed to near zero during training (due to the L1 regularization), leaving 465 active units. Inactive units were excluded from analysis and are not shown. Red – excitation, blue - inhibition. Inset shows axes.

Figure 4-Figure Supplement 6 | Full set of visual spatial RFs of temporal prediction

model units trained on visual inputs without added noise. Model units were obtained by training the model with 1600 hidden units on visual inputs. The hidden unit number and L1 weight regularization strength (10^{-4}) was chosen because it results in the lowest MSE on the prediction task, as measured using a cross validation set. Some hidden units' weight matrices decayed to near zero during training (due to the L1 regularization), leaving 1585 active units, which were included in analysis. Inactive units were excluded from analysis. Each square represents the spatial RF of a single unit, shown at its best time step. The best time step was determined by selecting the time step for which the power (sum of squares) of the RF was greatest. White – excitation, black - inhibition.

Figure 4-Figure Supplement 7 | 2D (space-time) visual spatiotemporal RFs of temporal prediction model units trained on visual inputs without added noise. Obtained from the same units shown in Fig. 4-Fig. Supplement 6 using methods outlined in Fig 2c. Red – excitation, blue - inhibition. Inset shows axes.

Figure 5-Figure Supplement 1 | Full set of visual spatial RFs of sparse coding model units. Model units were obtained by training the sparse coding model with 3200 units on identical visual inputs used to train the temporal prediction model. The model configuration (3200 units, L1 sparsity strength of $10^{0.5}$ on the unit activities) was chosen because it resulted in the RFs that look most like the RFs of V1 simple cells as determined by visual inspection. Each square represents the spatial RF of a single unit, shown at its best time step. The best time step was determined by selecting the time step for which the power (sum of squares) of the RF was greatest. White – excitation, black - inhibition.

Figure 5-Figure Supplement 2 | 2D (space-time) visual spatiotemporal RFs of sparse coding model units. Obtained from the same units shown in Fig. 5-Fig. Supplement 1 using methods outlined in Fig 2c. Red – excitation, blue - inhibition. Inset shows axes.

Figure 5-Figure Supplement 3 | Full set of auditory RFs of sparse coding model trained on auditory inputs without added noise. Units were obtained by training the sparse coding model with 1600 units on the identical auditory inputs used to train the network shown in Fig. 7-Fig. Supplement 2. L1 regularization of $10^{0.5}$ was applied to the units' activities. This network configuration was selected as it produced unit RFs that most closely resembled those recorded in A1, as determined by visual inspection. Red – excitation, blue - inhibition. Inset shows axes.

Figure 5-Figure Supplement 4 | Full set of visual spatial RFs of sparse coding model units trained on visual inputs without added noise. Model units were obtained by training the sparse coding model with 3200 units on identical visual inputs used to train the temporal prediction model. The model configuration (3200 units, L1 sparsity strength of $10^{0.5}$ on the unit activities) was chosen because it resulted in the RFs that look most like

the RFs of V1 simple cells as determined by visual inspection. Each square represents the spatial RF of a single unit, shown at its best time step. The best time step was determined by selecting the time step for which the power (sum of squares) of the RF was greatest. White – excitation, black - inhibition.

Figure 5-Figure Supplement 5 | 2D (space-time) visual spatiotemporal RFs of sparse coding model units trained on visual inputs without added noise. Obtained from the same units shown in Fig. 5-Fig. Supplement 4 using methods outlined in Fig 2c. Red – excitation, blue - inhibition. Inset shows axes.

Figure 6-Figure Supplement 1 | Population measures for real A1 spectrotemporal RFs and temporal prediction and sparse coding model auditory RFs when models are trained on auditory inputs without added noise. Real units are the same as those shown in Figure 3. Temporal prediction model units are the same as those shown in Fig. 4-Fig. Supplement 5; Sparse coding model units are the same as those shown in Fig. 5-Fig. Supplement 3. **a**, Each point represents a single RF (with 32 frequency and 38 time steps) which has been embedded in a 2 dimensional space using Multi-Dimensional Scaling (MDS). Red circles - real A1 neurons, black circles – temporal prediction model units, blue triangles – sparse coding model units. Colour scheme applies to all subsequent panels in Figure. **b**, Proportion of power contained in each time step of the RF, taken as an average across the population of units. **c**, Temporal span of excitatory subfields versus that of inhibitory subfields, for real neurons and temporal prediction and sparse coding model units. The area of each circle is proportional to the number of occurrences at that point. The inset plots, which zoom in on the distribution use a smaller constant of proportionality for the circles to make the distributions clearer. **d**, Distribution of temporal spans of excitatory subfields, taken by summing along the x-axis in **c**. **e**, Distribution of temporal spans of inhibitory subfields, taken by summing along the y-axis in **c**. **f**, Frequency span of excitatory subfields versus that of inhibitory subfields, for real neurons and temporal prediction and sparse coding model units. **g**, Distribution of frequency spans of excitatory subfields, taken by summing along the x-axis in **f**. **h**, Distribution of frequency spans of inhibitory subfields, taken by summing along the y-axis in **f**. The addition of noise leads to subtle changes in the RFs. Without noise, the inhibition in the temporal prediction model tends to be slightly less extended and the RFs a little less smooth (see Figure 4, Figure 4 - Figure supplement 5 for qualitative comparison).

Figure 7-Figure Supplement 1 | Visual RFs and population measures for real V1 neurons and sparse coding model units. **a**, Model units are the same as those used in Fig. 5-Fig. Supplement 1. Example spatial RFs of randomly selected units at their best time step. **b-c**, Example 3D and corresponding 2D spatiotemporal RFs at most recent 6 time steps of **(b)** (I, space-time separable, and II, space-time inseparable) real²³ V1 neurons and **(c)** (I-III, space-time separable, and IV-VI, space-time inseparable) sparse coding model units. **d**, Proportion of power (sum of squared weights over space and averaged across units) in each time step, for real and model populations. **e**, Joint

distribution of spatial frequency and orientation tuning for population of model units. **f**, Distribution of RF shapes for real neurons (cat⁴⁰, mouse⁵⁹ and monkey⁴²) and model units. For **e-f**, only units that could be well approximated by Gabor functions (n = 2402 units; see Methods) were included in the analysis. Of these, only model units that were space-time separable (n = 881) are shown in **f** to be comparable with the neuronal data¹⁶.

Figure 7-Figure Supplement 2 | Visual RFs and population measures for real V1 neurons and temporal prediction model units trained on visual inputs without added noise. **a**, Model units are the same as those used in Fig. 11. Example spatial RFs of randomly selected units at their best time step. **b-c**, Example 3D and corresponding 2D spatiotemporal RFs at most recent 6 time steps of (I, space-time separable, and II, space-time inseparable) real²³ V1 neurons and (c) (I-III, space-time separable, and IV-VI, space-time inseparable) sparse coding model units. **d**, Proportion of power (sum of squared weights over space and averaged across units) in each time step, for real and model populations. **e**, Joint distribution of spatial frequency and orientation tuning for population of model units. **f**, Distribution of orientation tuning for population of model units. **g**, Distribution of RF shapes for real neurons (cat⁴⁰, mouse⁵⁹ and monkey⁴²) and model units. For **e-g**, only units that could be well approximated by Gabor functions (n = 1246 units; see Methods) were included in the analysis. Of these, only model units that were space-time separable (n = 569) are shown in **f** to be comparable with the neuronal data¹⁶. The addition of noise only leads to subtle changes in the RFs; most apparently, there are more units with RFs comprising multiple short subfields (forming an increased number of points towards the lower right quadrant of **g**) than is seen in the case when noise is used.

Figure 7-Figure Supplement 3 | Visual RFs and population measures for real V1 neurons and sparse coding model units trained on visual inputs without added noise. **a**, Model units are the same as those used in Fig. 15. Example spatial RFs of randomly selected units at their best time step. **b-c**, Example 3D and corresponding 2D spatiotemporal RFs at most recent 6 time steps of (b) (I, space-time separable, and II, space-time inseparable) real²³ V1 neurons and (c) (I-III, space-time separable, and IV-VI, space-time inseparable) sparse coding model units. **d**, Proportion of power (sum of squared weights over space and averaged across units) in each time step, for real and model populations. **e**, Joint distribution of spatial frequency and orientation tuning for population of model units. **f**, Distribution of orientation tuning for population of model units. **g**, Distribution of RF shapes for real neurons (cat⁴⁰, mouse⁵⁹ and monkey⁴²) and model units. For **e-g**, only units that could be well approximated by Gabor functions (n = 2482 units; see Methods) were included in the analysis. Of these, only model units that were space-time separable (n = 860) are shown in **f** to be comparable with the neuronal data¹⁶.

Figure 8-Figure Supplement 1 | Correspondence between sparse coding model's ability to reproduce its input and the similarity of its units' responses to those of real A1 neurons. Performance of model as a function of number of units and regularization on the weights as measured by **a**, reconstruction error on validation set at

the end of training and **b**, similarity between model units and real A1 neurons. The similarity between the real and model units is measured by averaging Kolmogorov-Smirnov distance between each of the real and model distributions for the span of temporal and frequency tuning of the excitatory and inhibitory RF subfields (e.g. the distributions in Fig. 3d-e and Fig. 3g-h).

Figure 8-Figure Supplement 2 | Interactive figure exploring the relationship between the strength of L1 regularisation on the network weights and the structure of the RFs the network produces when the network is trained on auditory inputs.

The interactive version of this figure can be found at:

https://yossing.github.io/temporal_prediction_model/figures/interactive_supplementary_figures.html

The left hand panel shows the performance of the network with the hyperparameter settings specified on the x and y axes. The x axis signifies the strength of L1 regularisation placed on the weights of the network during training. The y axis signifies the number of hidden units in the network. The colour represents the predictive capacity of the model as measured by the reconstruction error on a held out validation set.

How to interact with the figure:

Hover over a point in the left hand panel to show the corresponding spectrotemporal receptive fields of the network in the right hand panel. Using the settings near the right hand panel, zoom, pan and reset the image to explore the shapes of the spectrotemporal receptive fields. Many hidden units' weight matrices decayed to near zero during training. Inactive units were excluded from analysis and are not shown.

Figure 8-Figure Supplement 3 | Interactive figure exploring the relationship between the strength of L1 regularisation on the network weights and the structure of the RFs the network produces when the network is trained on visual inputs.

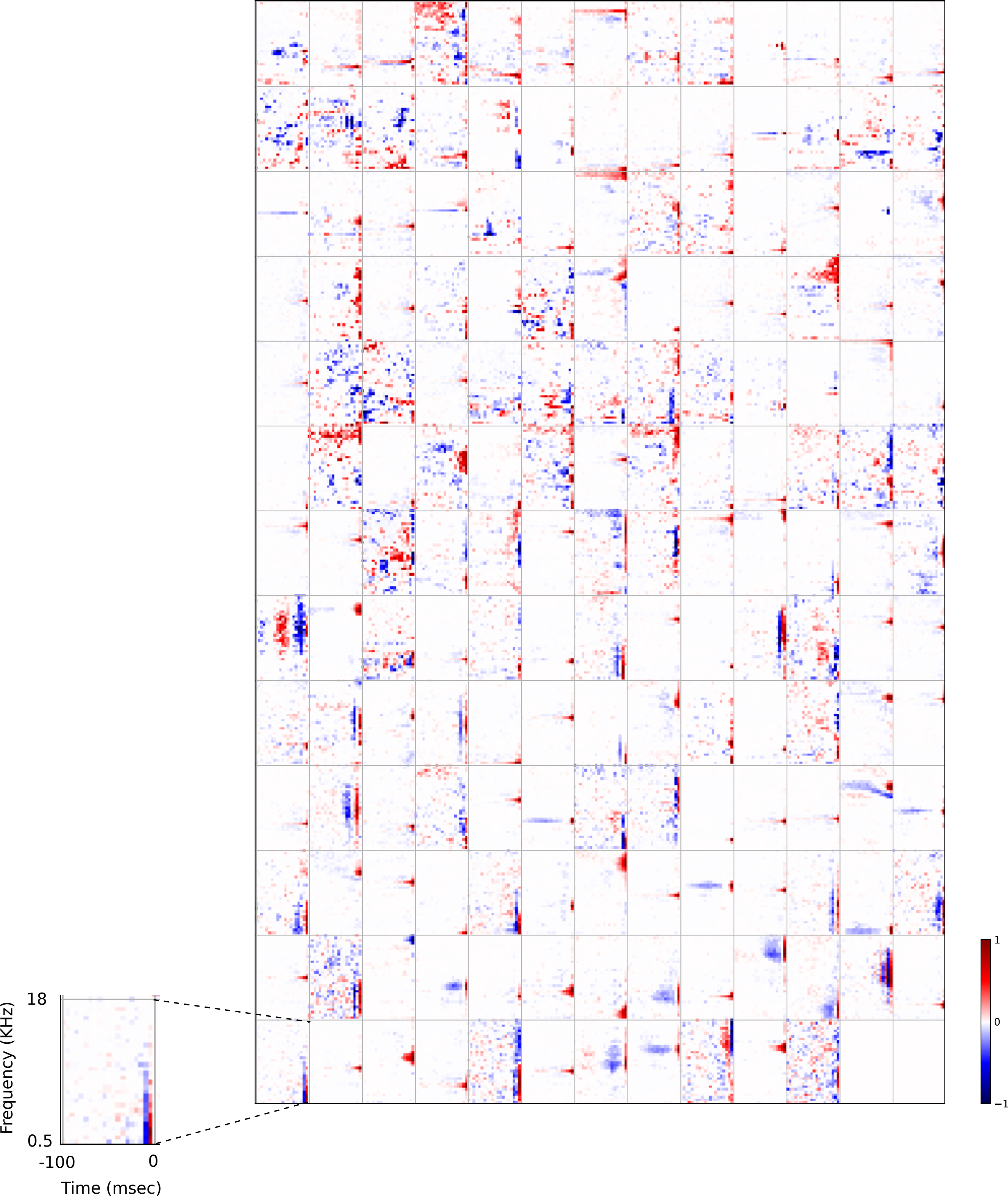
The interactive version of this figure can be found at:

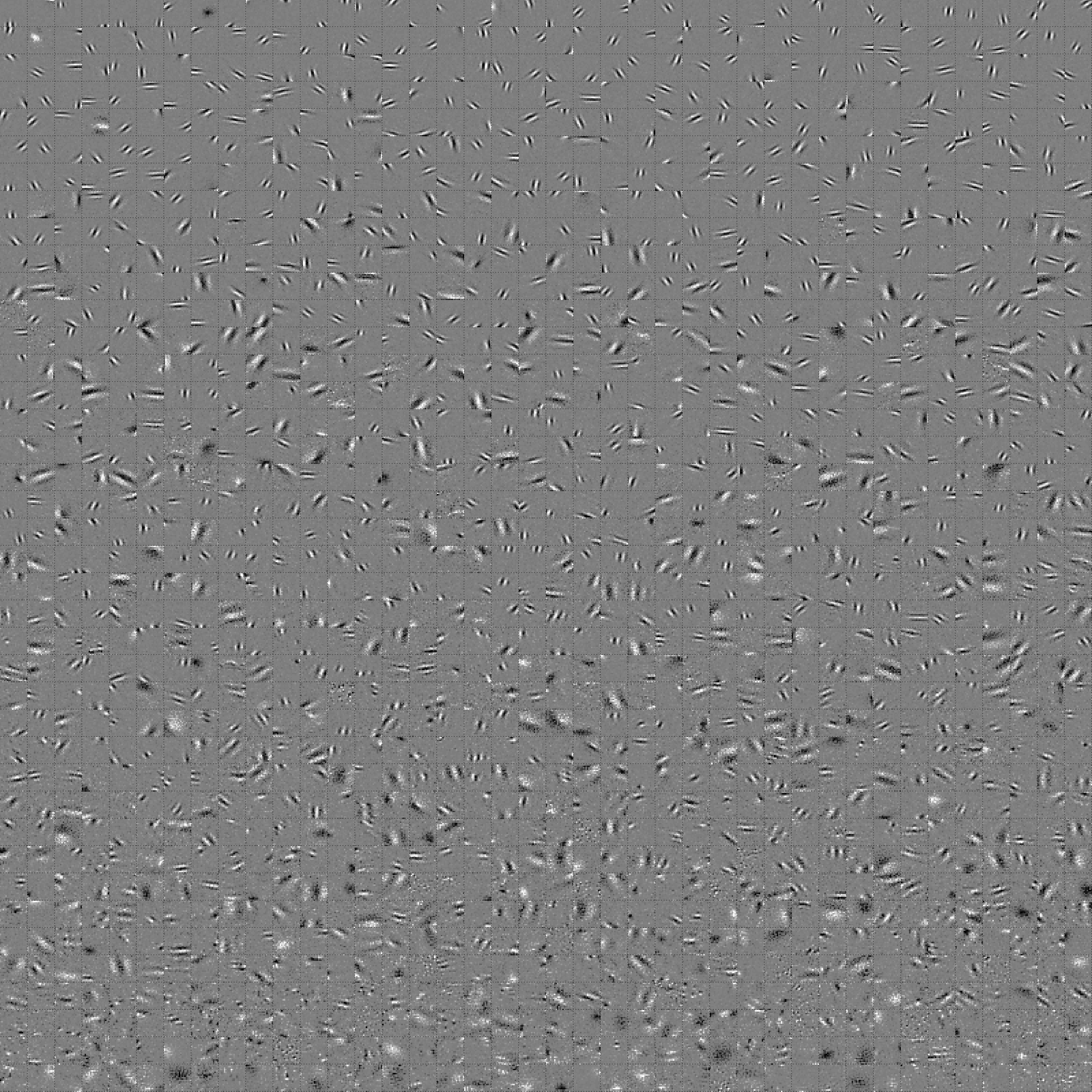
https://yossing.github.io/temporal_prediction_model/figures/interactive_supplementary_figures.html

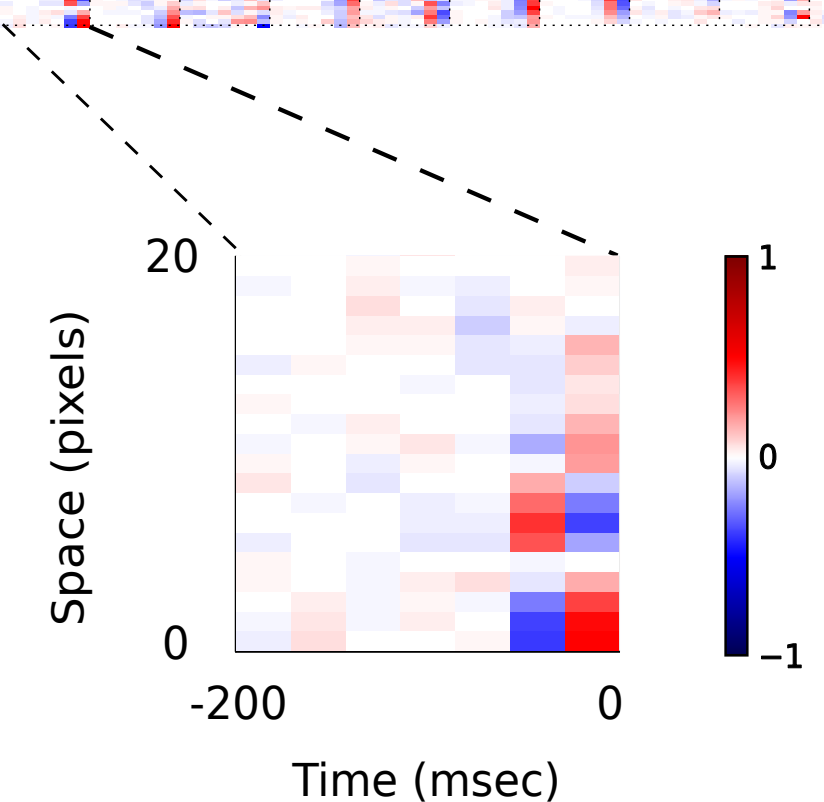
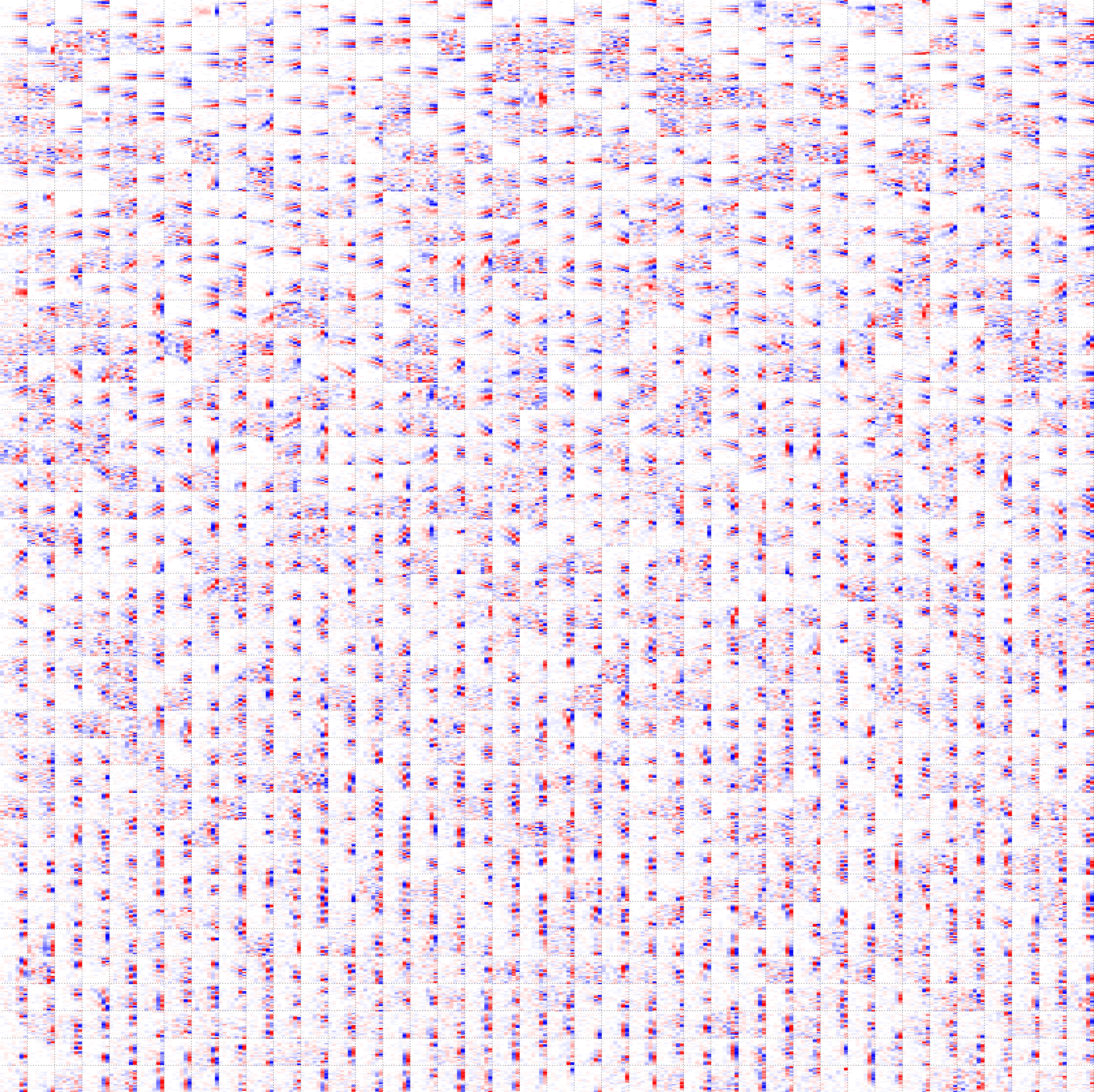
The left panel shows the performance of the network with the hyperparameter settings specified on the x and y axes. The x axis signifies the strength of L1 regularisation placed on the weights of the network during training. The y axis signifies the number of hidden units in the network. The colour represents the predictive capacity of the model as measured by the reconstruction error on a held out validation set.

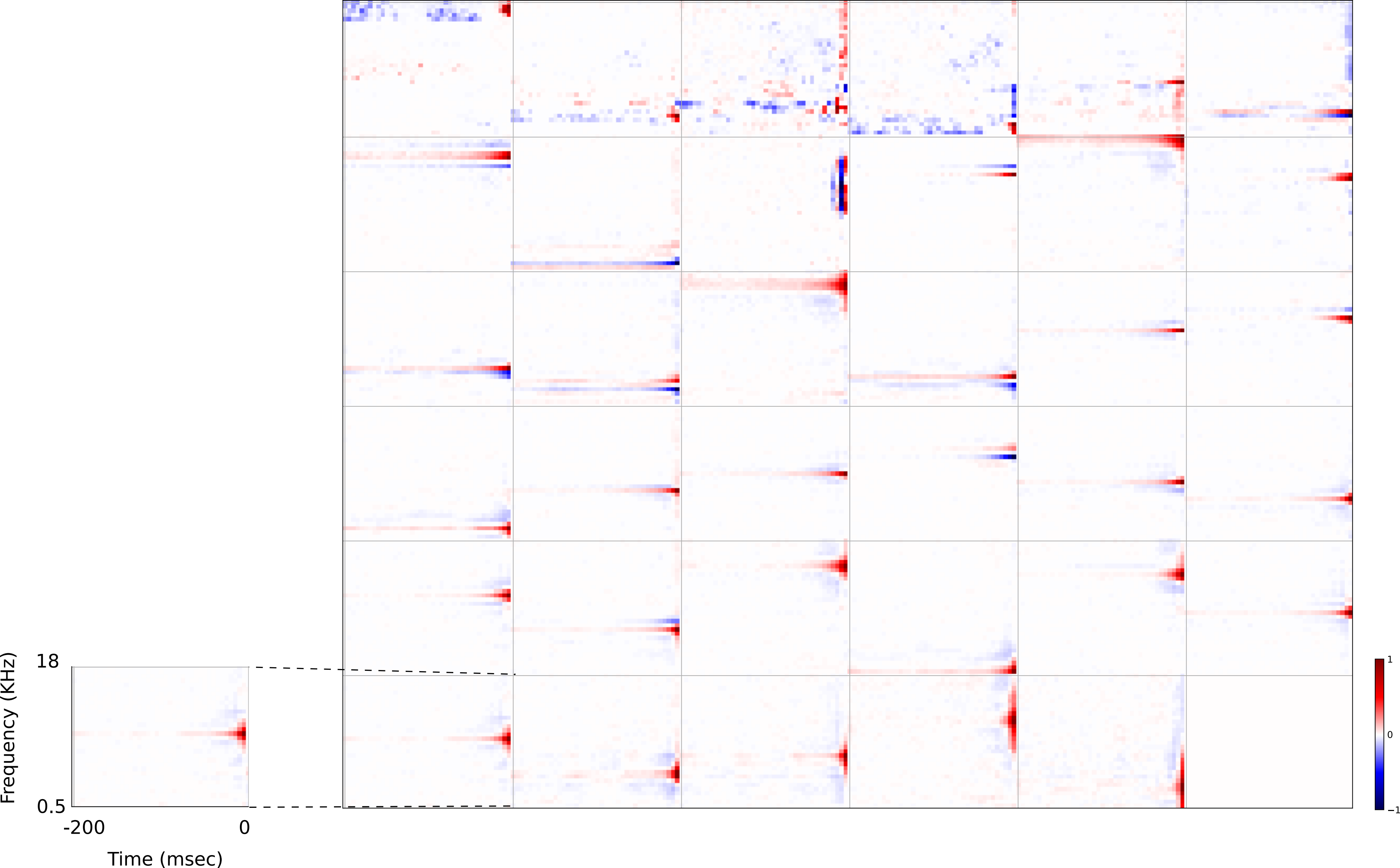
How to interact with the figure:

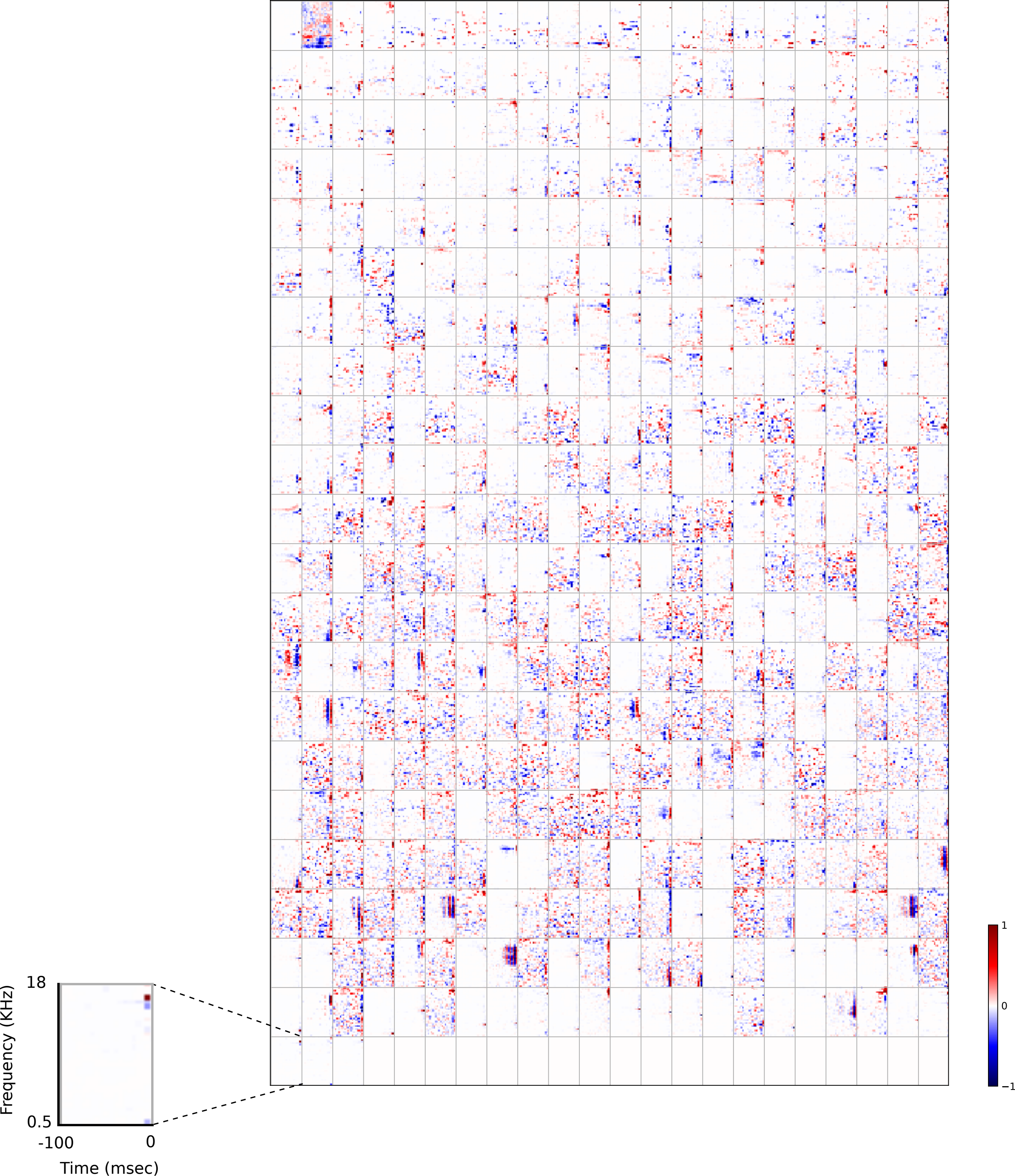
1432 Hover over a point in the left panel to show the corresponding spatial receptive fields of the
1433 network in the right panel. Using the settings on the right of the right hand panel, zoom,
1434 pan and reset the image to explore the shapes of the spatial receptive fields. Change the
1435 slider labelled 'time step' to change the time-step of the spatial receptive fields being
1436 shown in the right hand panel. Some hidden units' weight matrices decayed to near zero
1437 during training. Inactive units were excluded from analysis and are not shown
1438

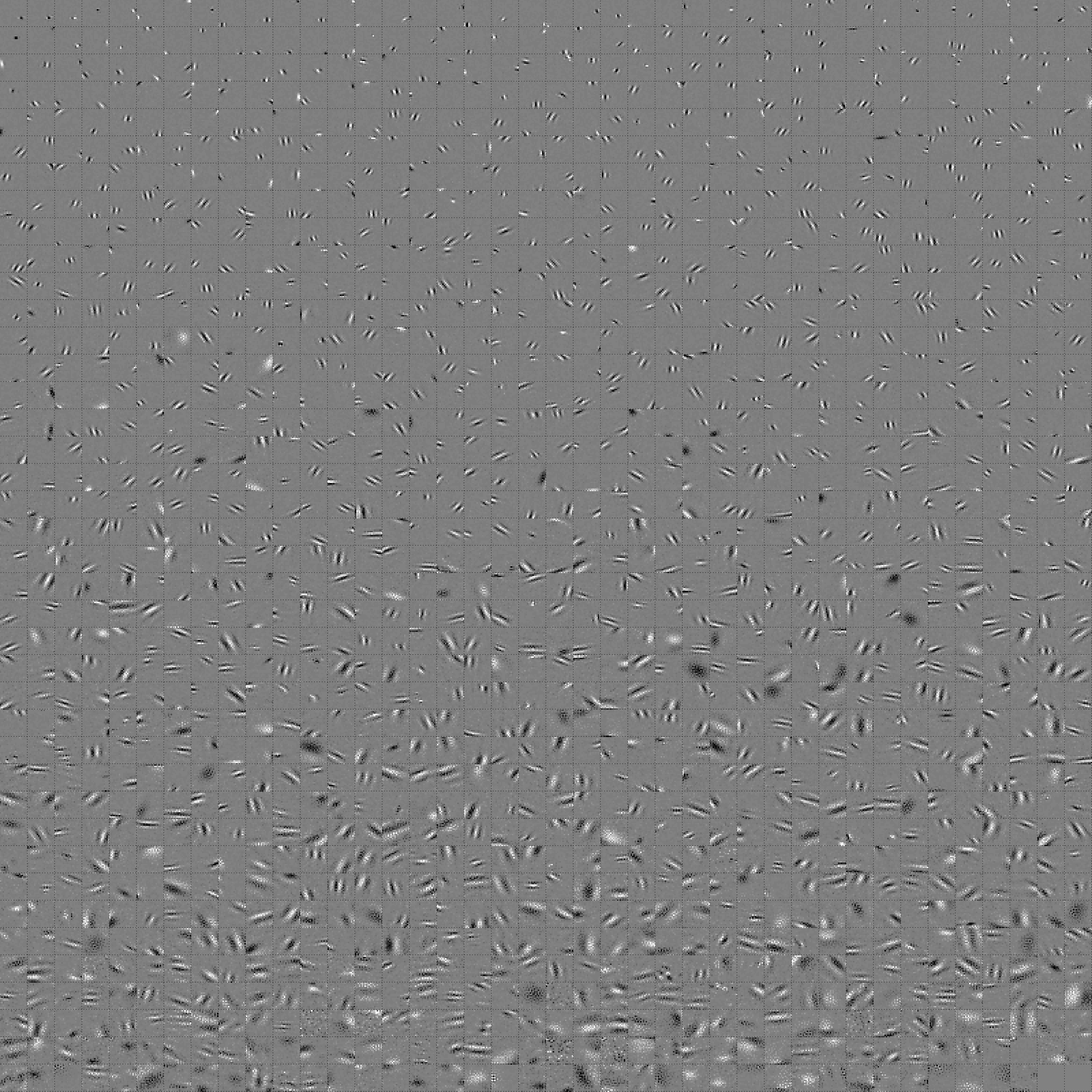


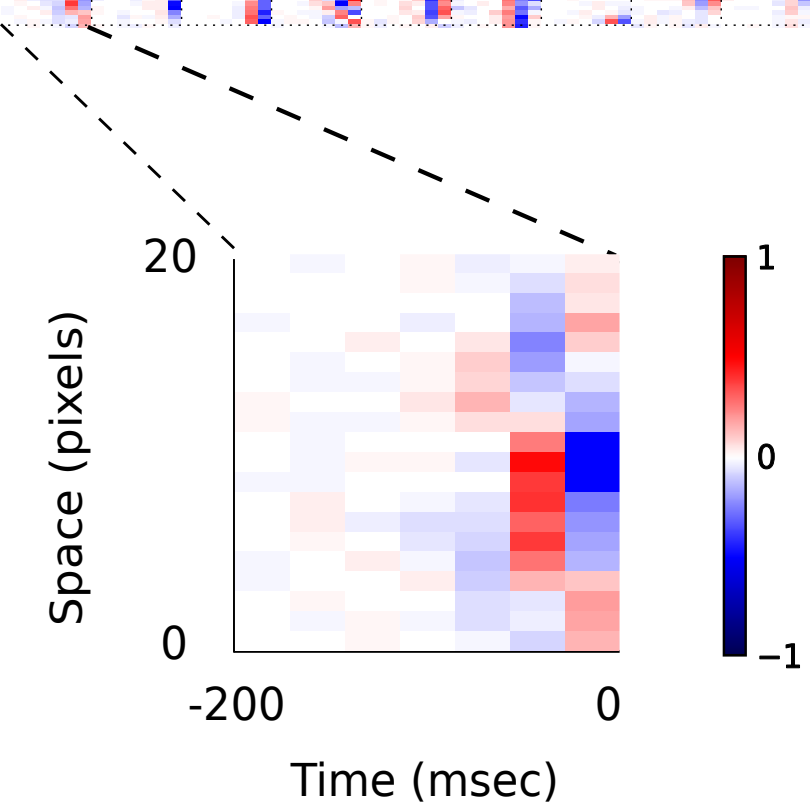
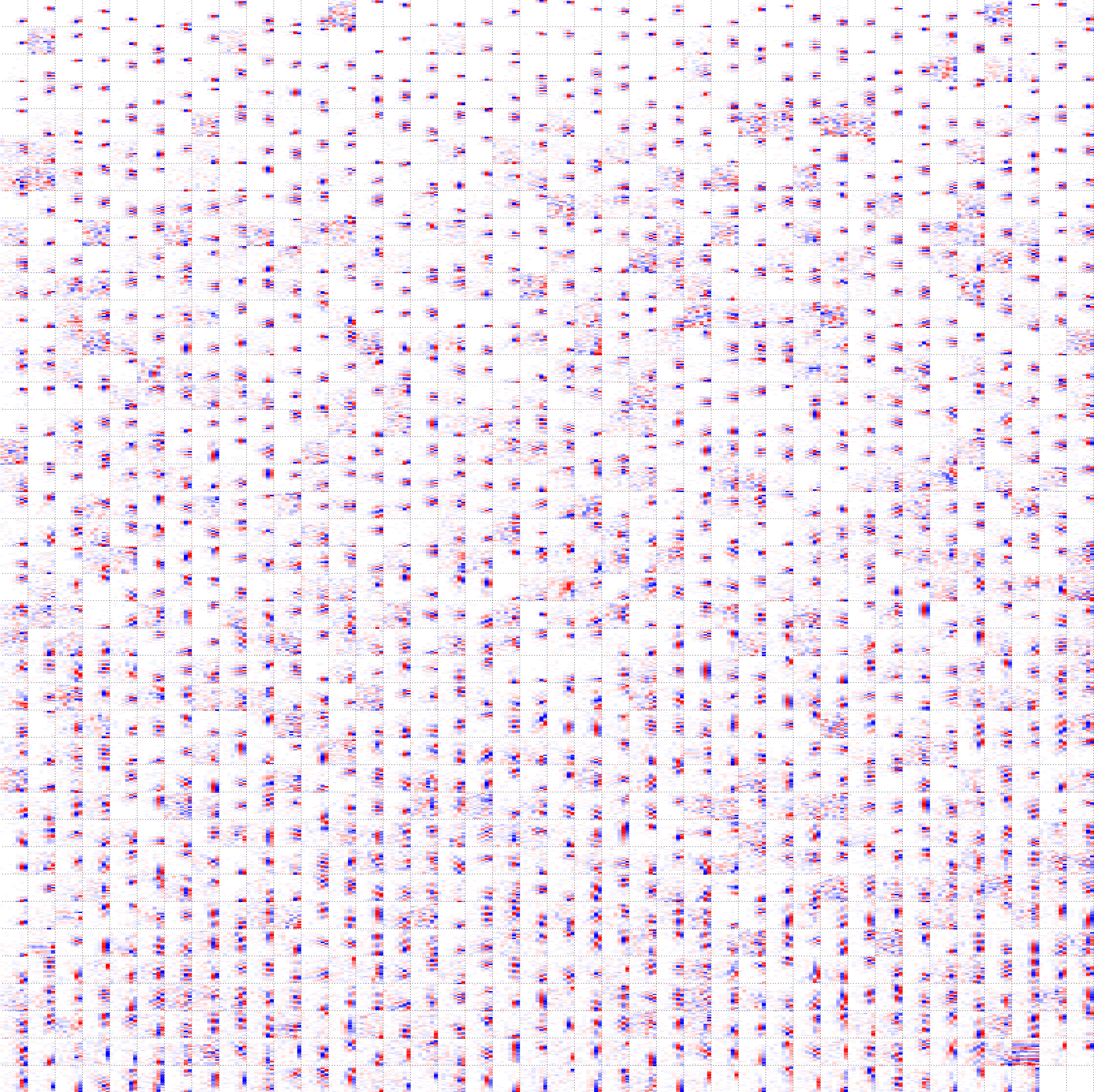


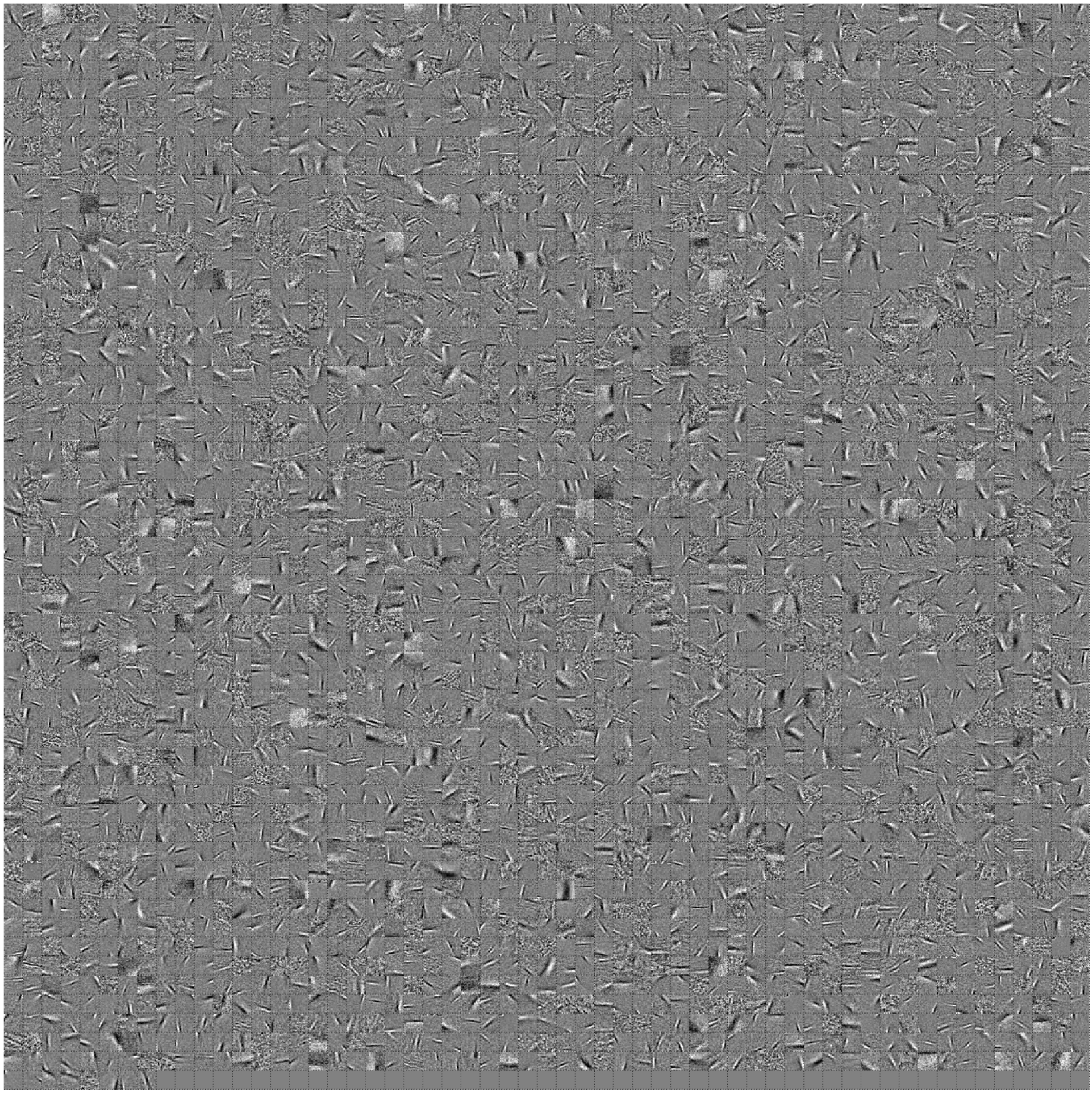


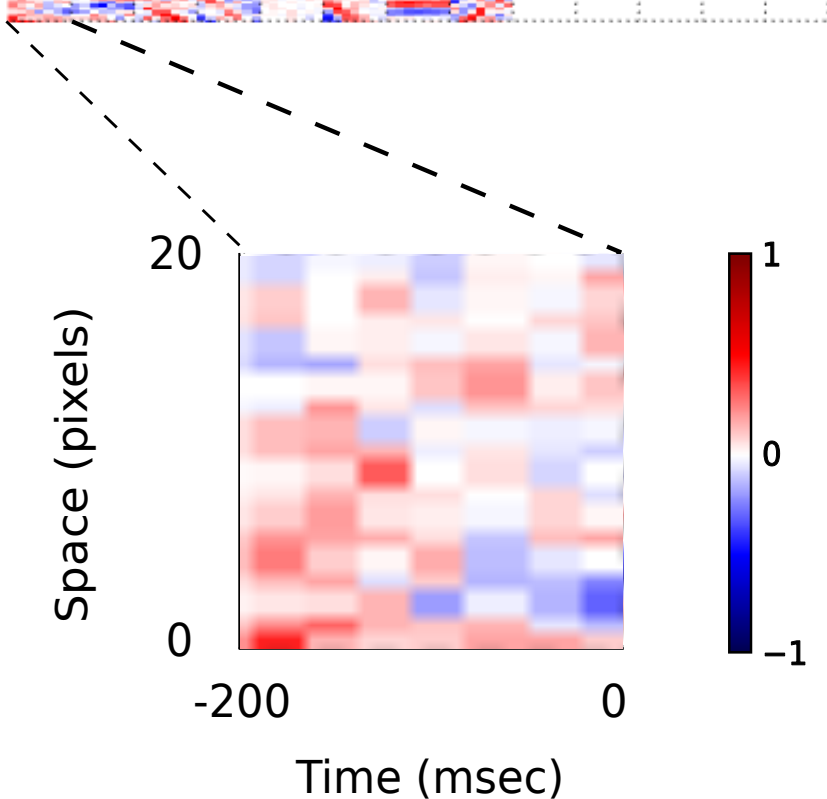
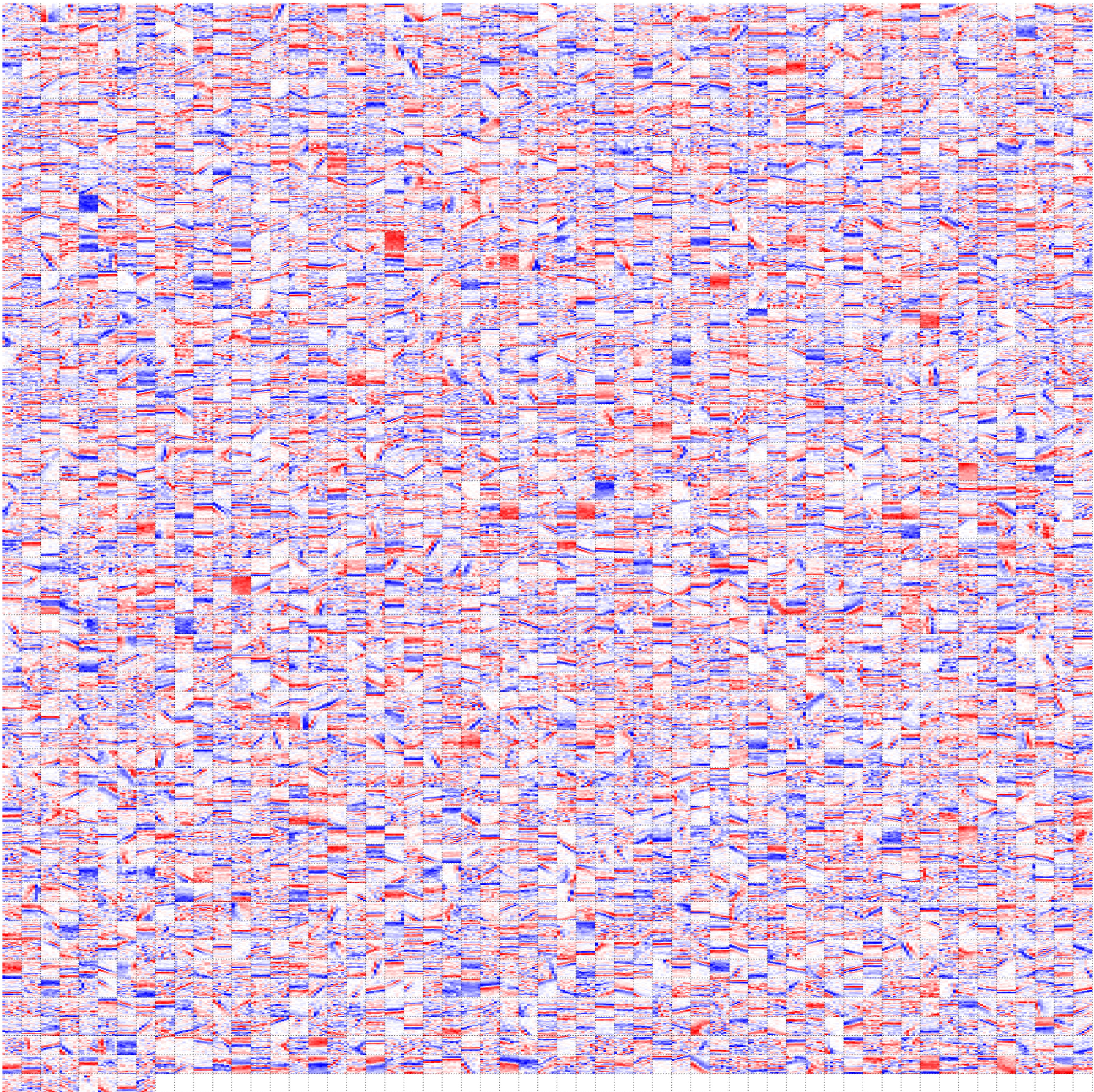


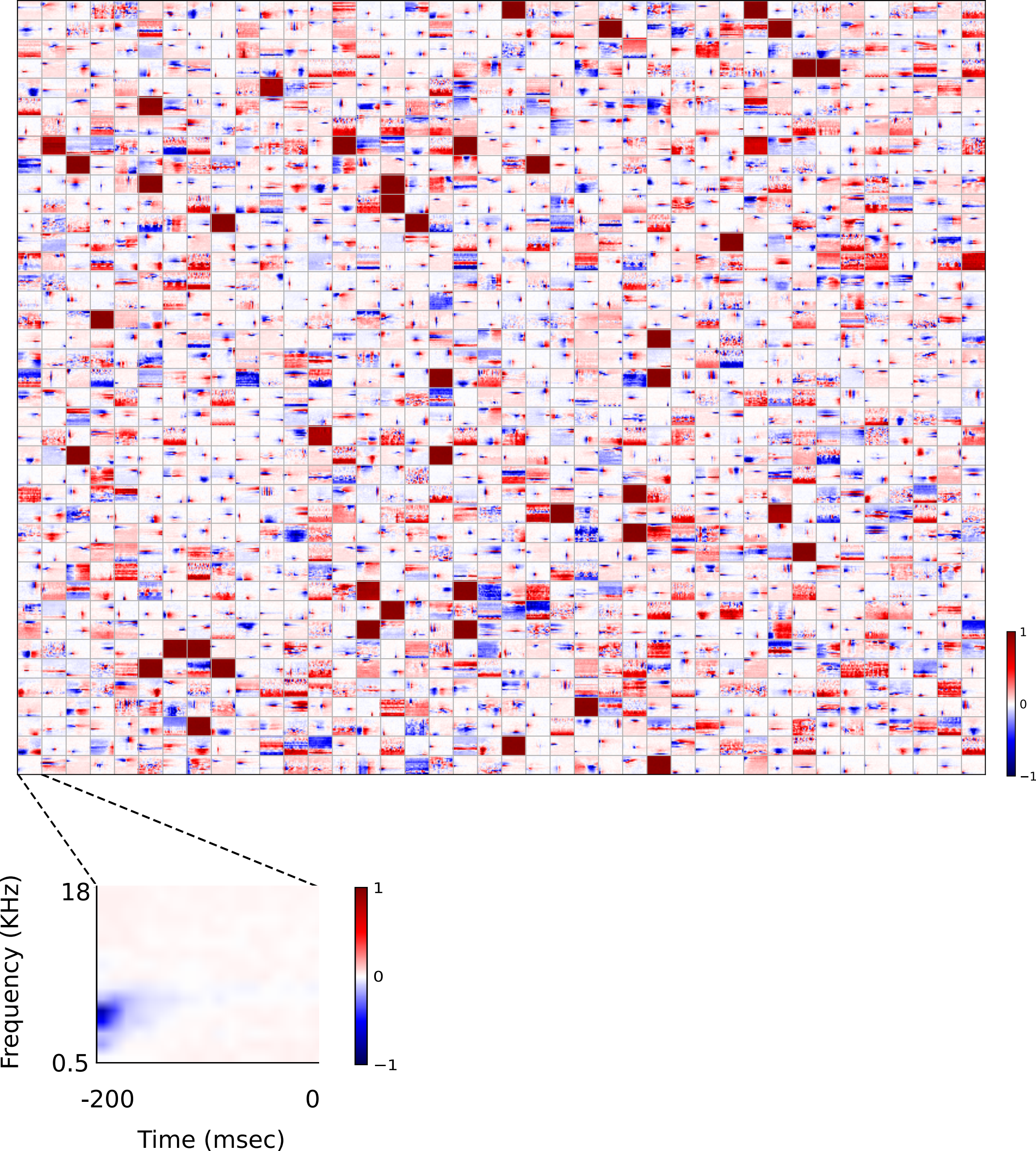


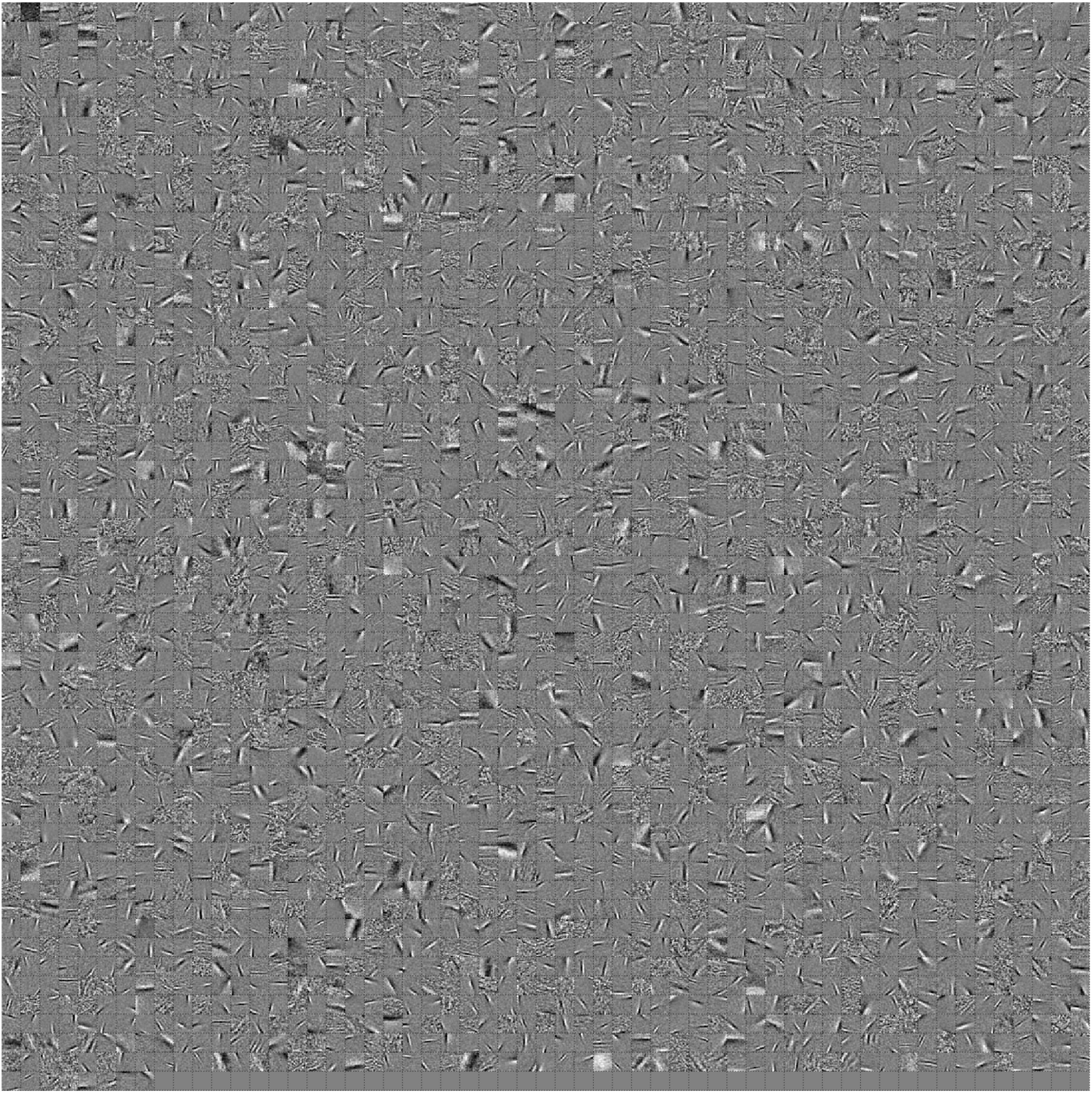


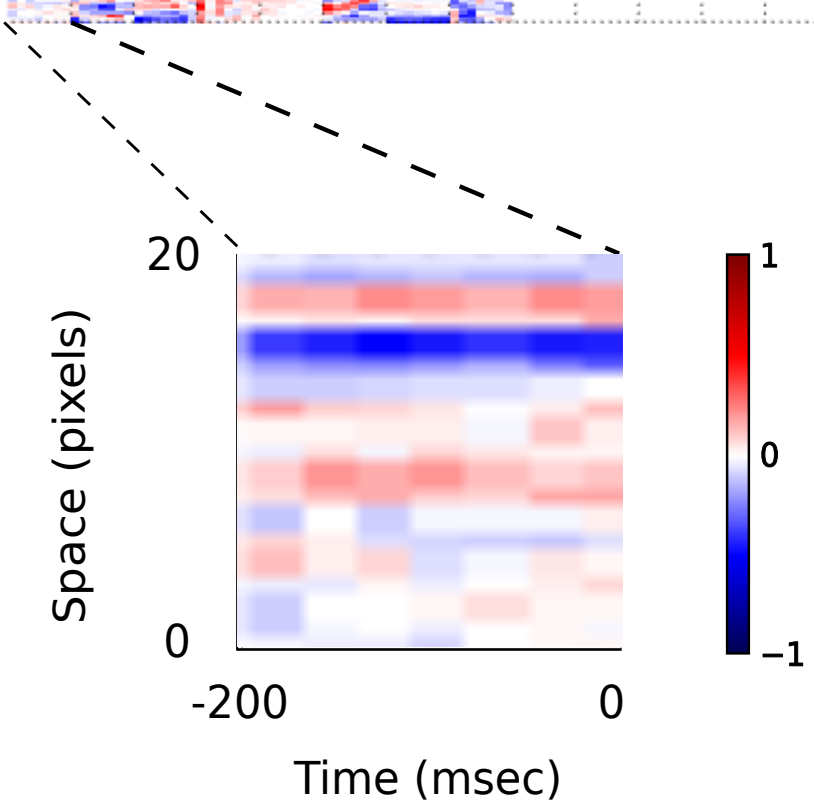
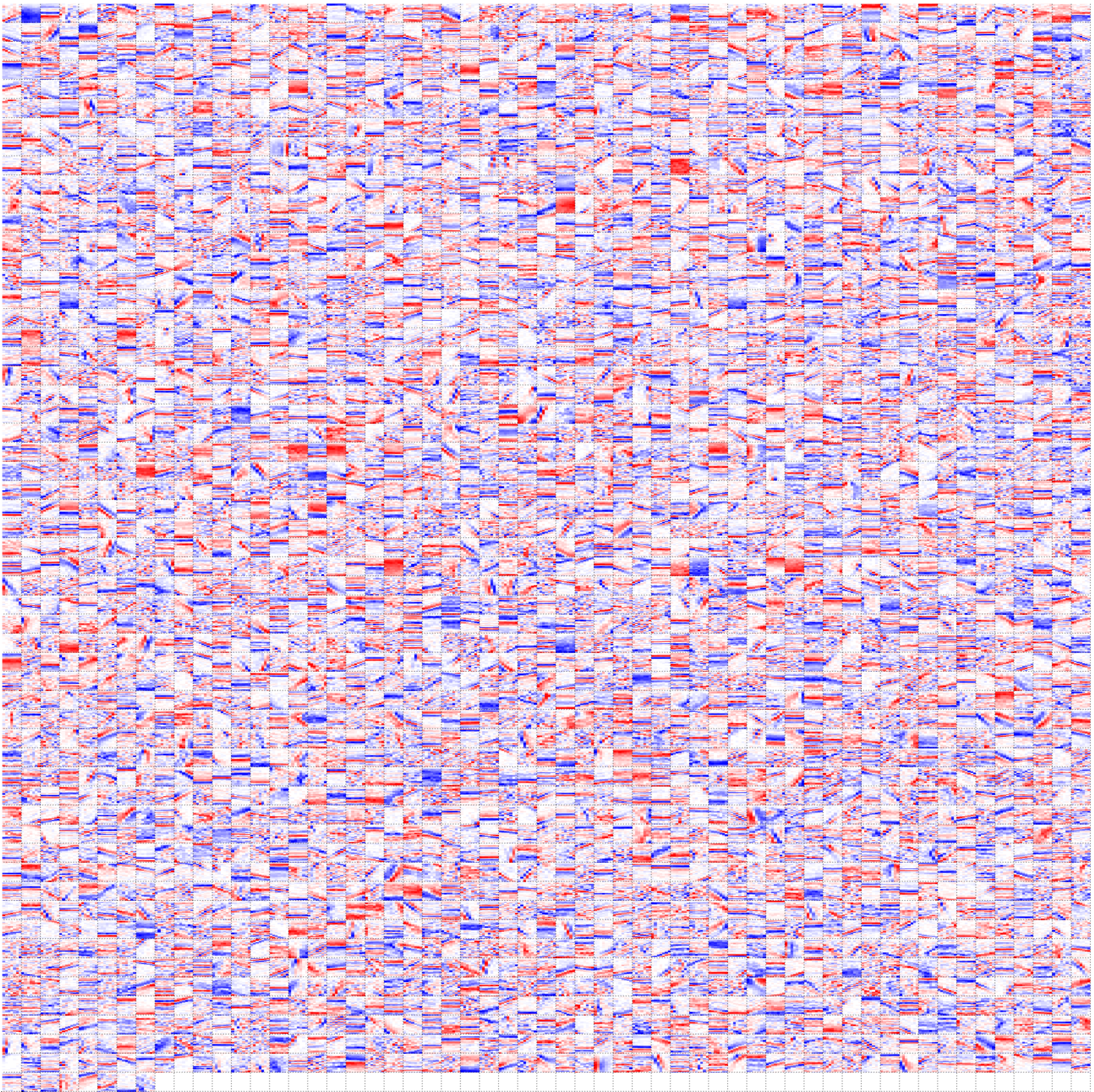


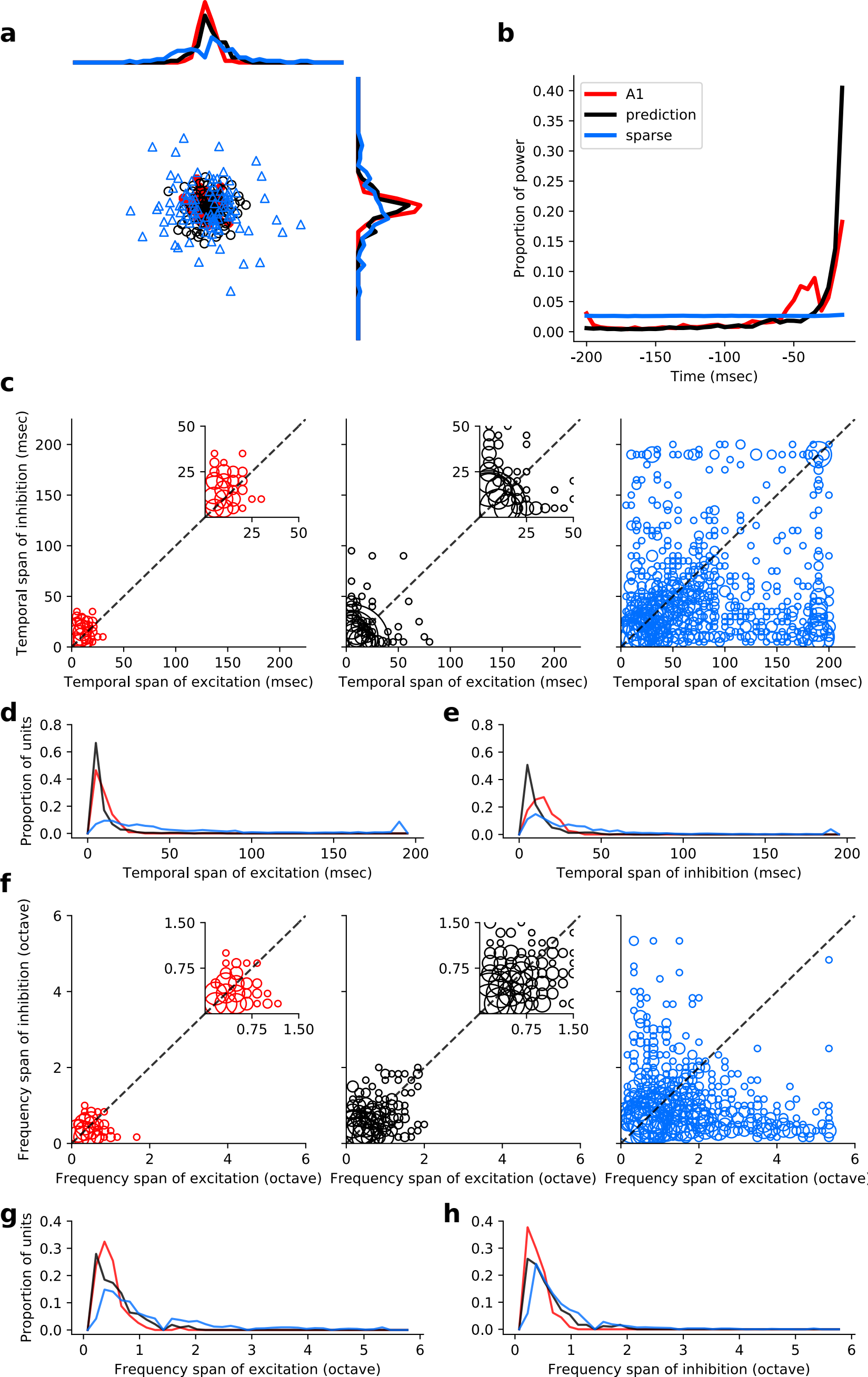


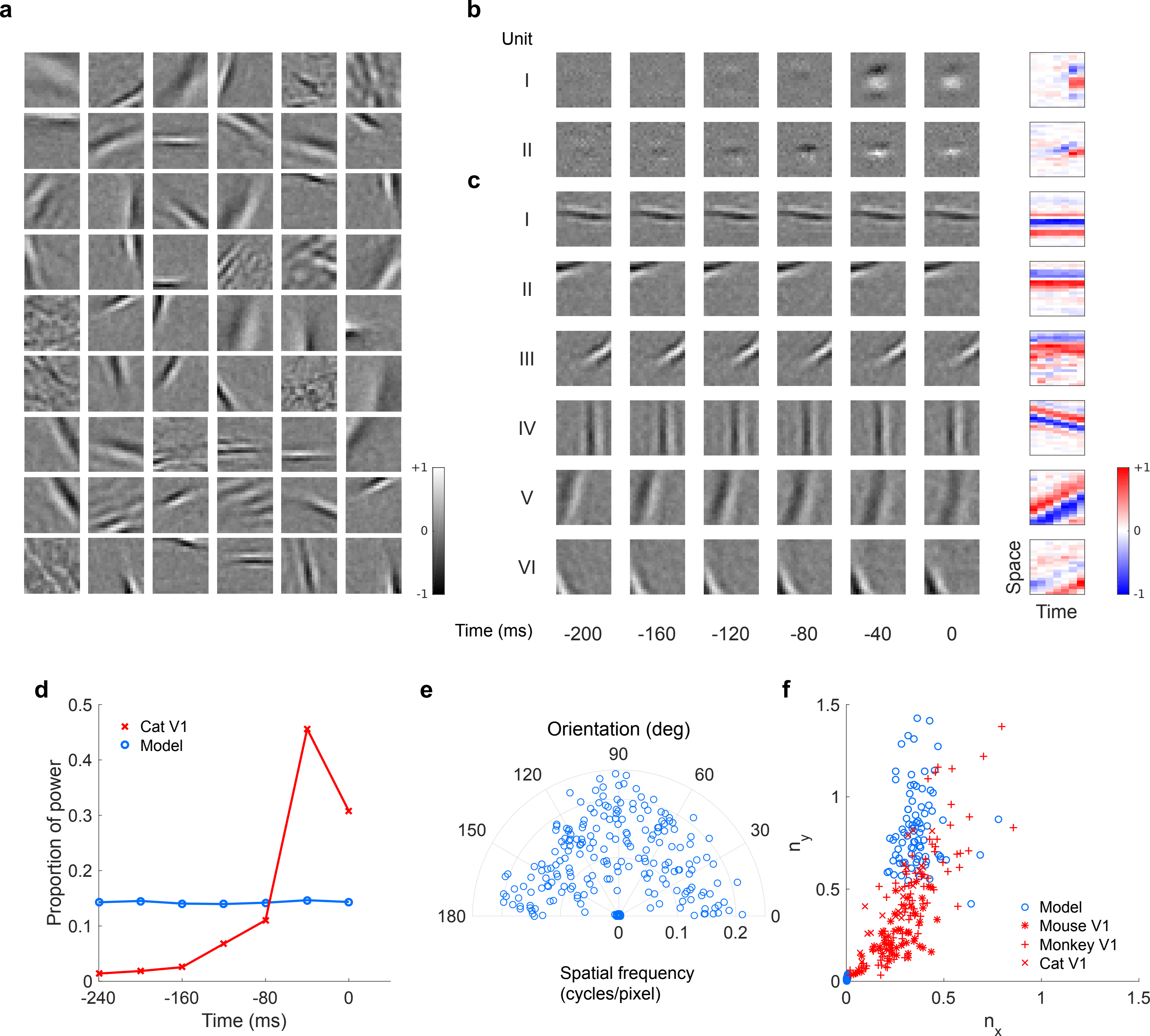


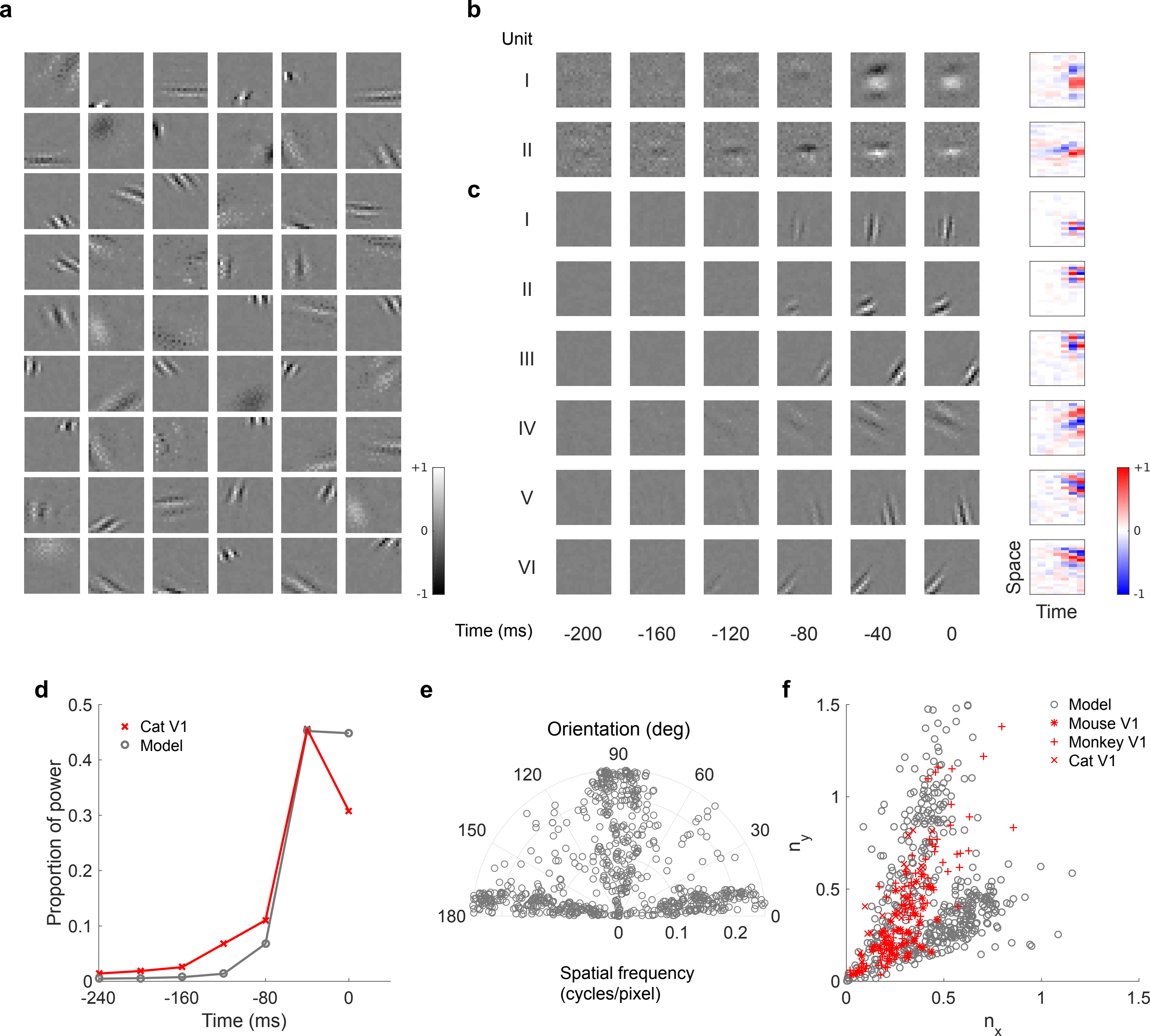


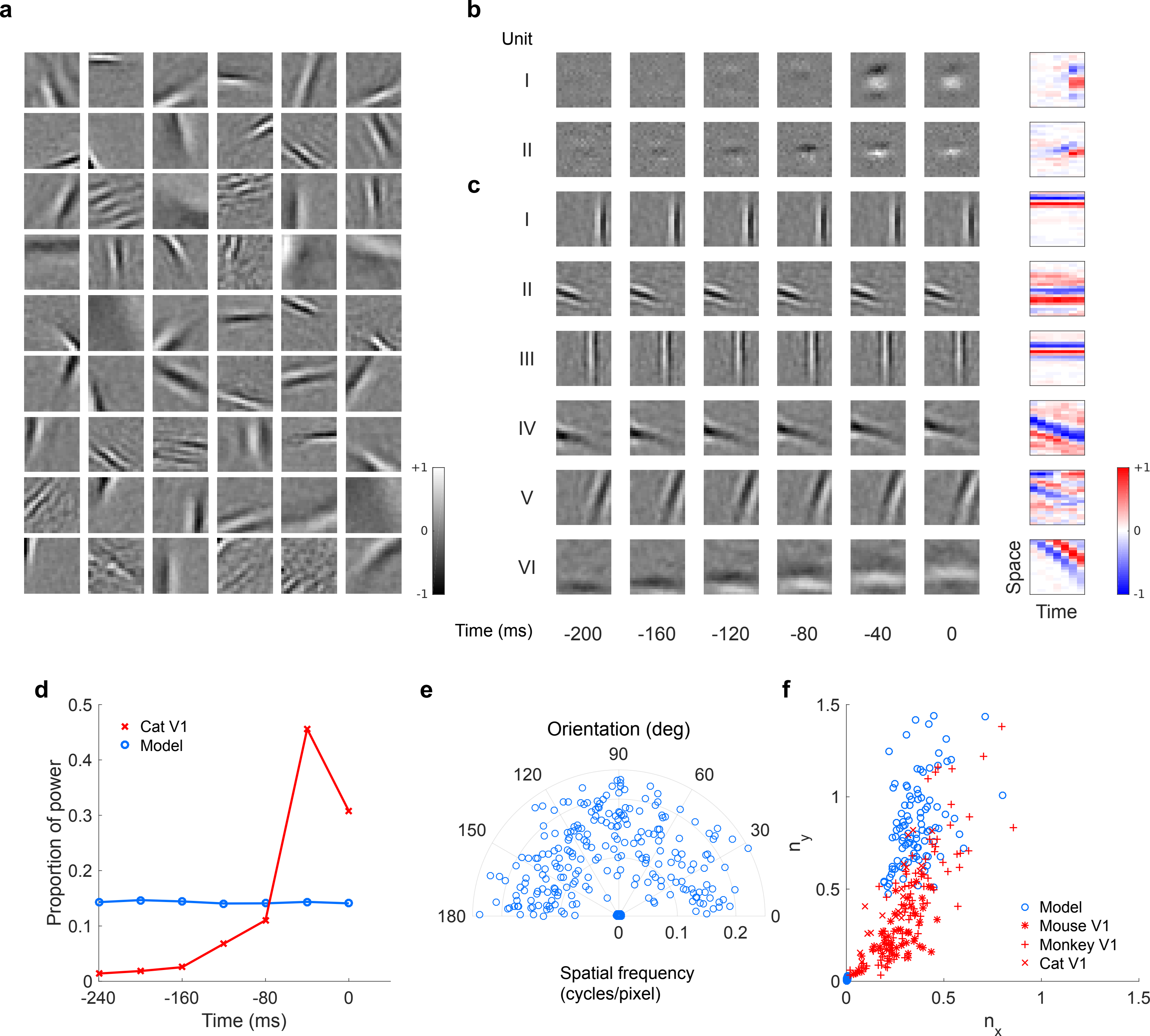


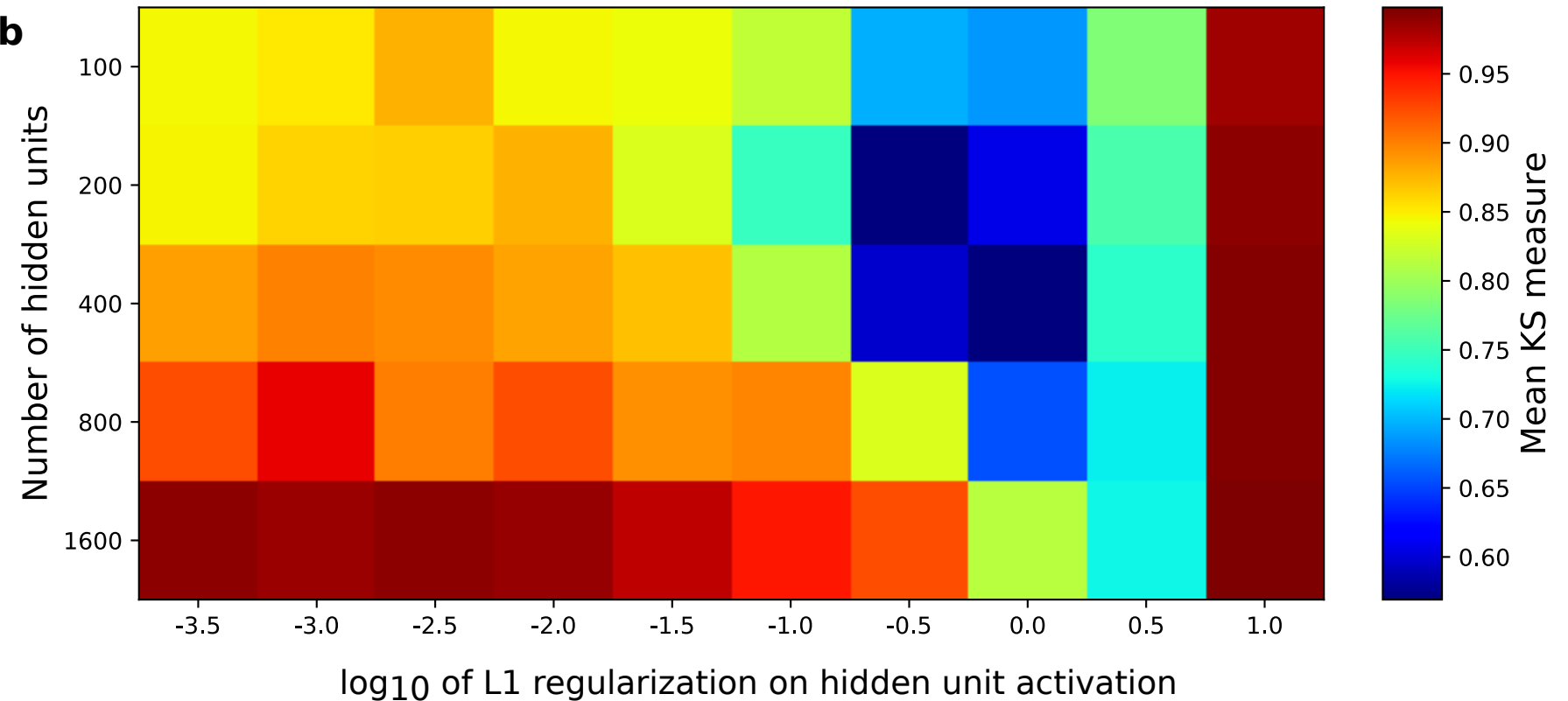
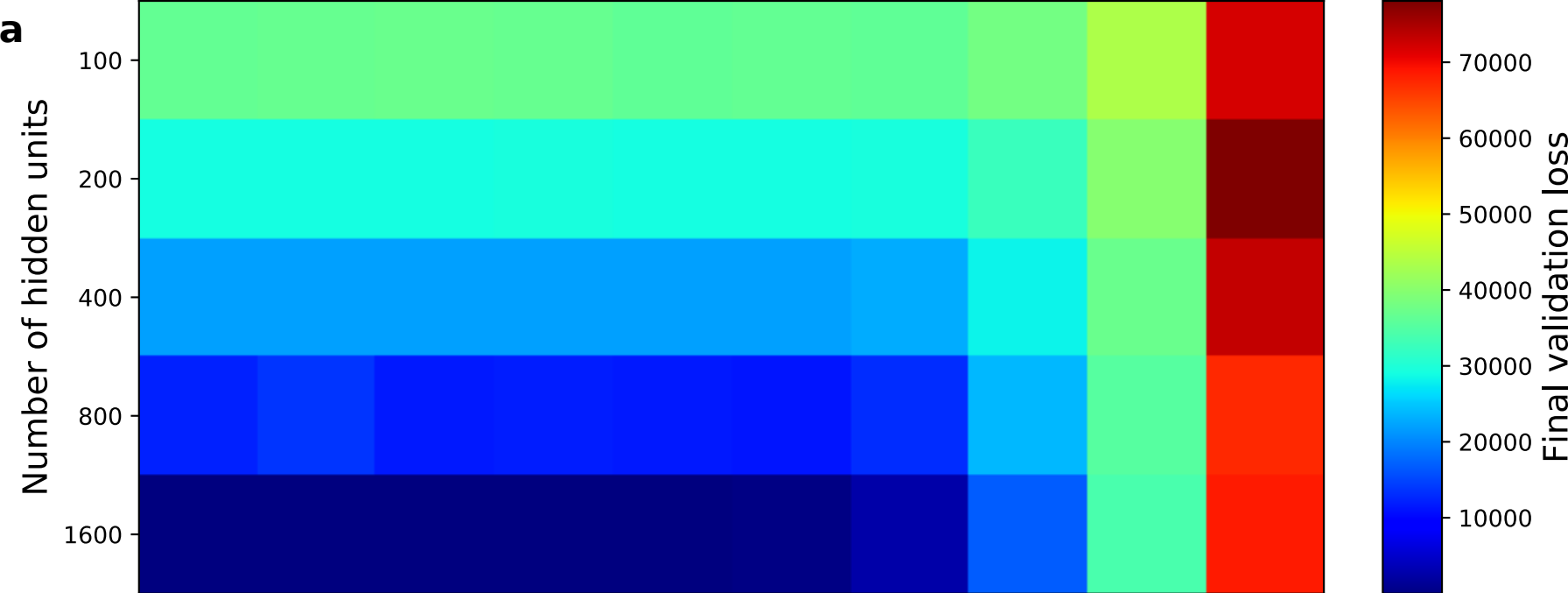




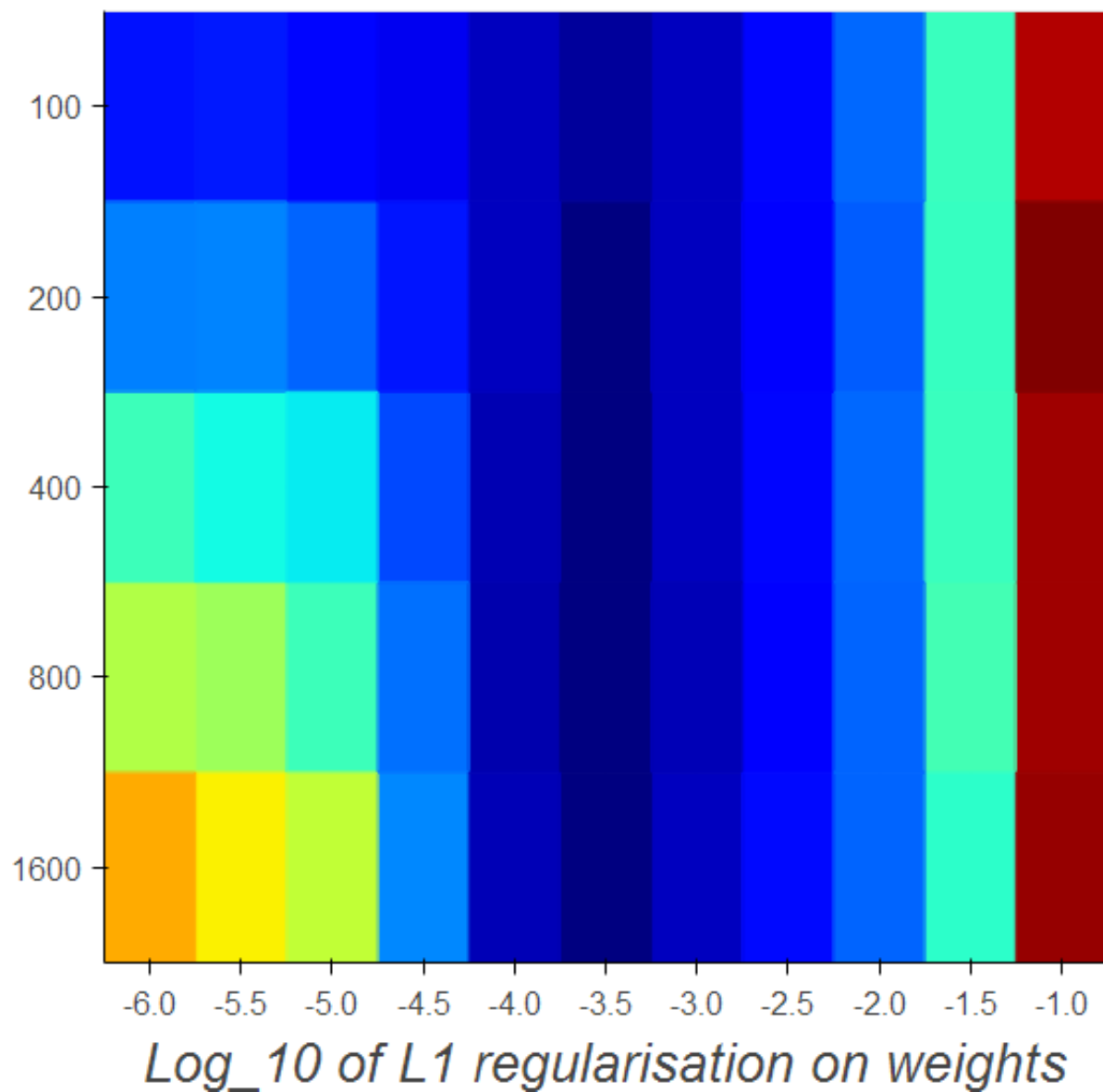








Heatmap



Corresponding spectrotemporal RFs

