Decoding movie content from neuronal population activity in the human medial temporal lobe

Franziska Gerken [1*], Alana Darcher [2*], Pedro J Gonçalves [3,4,5], Rachel Rapp [3], Ismail Elezi [1§], Johannes Niediek, [2, 8], Marcel S Kehl [2, 9], Thomas P Reber [2, 10], Stefanie Liebe, [3,6], Jakob H Macke [3,7‡], Florian Mormann [2‡], Laura Leal-Taixé [1†‡]

[1]Dynamic Vision and Learning Group, Technical University of Munich, Munich, Germany; [2]Dept. of Epileptology, University Medical Center of Bonn, Bonn, Germany; [3]Machine Learning in Science, Excellence Cluster Machine Learning and Tübingen AI Center, University of Tübingen, Tübingen, Germany; [4]VIB-Neuroelectronics Research Flanders (NERF), Belgium; [5]imec, Belgium; [6]Neurology, University Hospital Tübingen, Tübingen, Germany; [7]Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany; [8]Machine Learning Group, Technische Universität Berlin, Berlin, Germany; [9]Department of Experimental Psychology, University of Oxford, Oxford, UK; [10]Faculty of Psychology, UniDistance Suisse, Brig, Switzerland

---

Abstract Neurons of the medial temporal lobe (MTL) form the basis of semantic representation in the human brain. While known to contain category-selective cells, it is unclear how the MTL processes naturalistic, dynamic stimuli. We studied 2286 neurons recorded from the hippocampus, parahippocampal cortex, amygdala, and entorhinal cortex of 29 intracranially-implanted patients during a full-length movie. While few neurons responded preferentially to semantic features, we could reliably predict the presence of characters, location, and visual transitions from the neuronal populations using a recurrent neural network. We show that decoding performance differs across regions based on the feature category, and that the performance is driven by feature-selective single neurons when decoding visual transitions such as camera cuts. These findings suggest that semantic representation in the MTL varies based on semantic category, with decoding information embedded in specific subsets of neurons for event-related features or distributed across the entire population for character and location-related features.

---

[*]These authors contributed equally to this work.
[‡]These authors contributed equally to this work.
[†]Currently at NVIDIA.
[§]Currently at Huawei.

For correspondence: florian.mormann@ukbonn.de (Florian Mormann); llealtaixe@nvidia.com (Laura Leal-Taixé)

Introduction

The medial temporal lobe (MTL) is a key area in the extraction and assembly of semantic representations from sensory inputs. Previous work has found populations of MTL cells which represent information on the basis of categories [1–3], e.g., discriminate faces from objects [4], as well as individual cells that code for the specific identity of a stimulus [5]. The latter, known as concept cells, are defined by their semantically invariant tuning; a single cell will selectively respond to a stimulus depicting the same object, regardless of angle, color, or sensory modality. Current evidence suggests that the tuning of concept cells may differ across regions of the MTL on a categorical basis, with the amygdala preferentially responding to living things [6,7], and the parahippocampal cortex preferentially responding to landscapes and scenes [8]. Furthermore, evidence from single neurons and local population activity shows that novel semantic associations can form within minutes [9–11], and that MTL neurons can maintain stable representations for several hours or days [12].

Much of our current understanding of semantic representations in the brain, and specifically concept cells, has been established from static presentations of semantic information, and it is currently unclear to what extent these insights translate to natural environments. Previous work using video stimuli has revealed the presence of cells sensitive to event boundaries [13–16], stimulus-driven inter-subject correlations of neural activity [17,18], single-trial signatures of movie content [16], and character-related population activity [19]. While such work has partially addressed which semantic features of a naturalistic, continuous stimulus can be reliably decoded from the population activity of single neurons within the human MTL, there is presently no direct comparison between the representation of semantic information on the single-neuron level versus the population level under dynamic viewing conditions. As such, it is yet unknown whether the population responses underlying successful decoding [18,20,21], or encoding [22,23], of semantic categories is driven by a small number of semantically-tuned single neurons, relies on relevant information distributed throughout the neuronal population, or some combination therein.

In this work, we provide a machine learning-based decoding pipeline for analyzing the representation of semantic categories in single-neuron activity in the human medial temporal lobe during the presentation of a full-length commercial film. Our work addresses the following questions: i) which semantic features of a movie are represented by neuronal activity in the MTL, ii) are there region-specific patterns of decoding performance for these features, and iii) how is the relevant information distributed in the population of neurons? Using a recurrent neural network, we predicted character presence, setting, and visual transitions of the movie from the activity across recorded brain regions and from each region separately. While we successfully decoded all categories from populations of neurons across the MTL, individual neurons showed consistent stimulus-related changes in firing rate only for visual transitions (also known as event boundaries). To understand which aspect of the neuronal population activity influenced decoding, we performed a series of follow-up analyses: First, we found that single neurons with stimulus-related responses accounted for the prediction performance on the population level for visual transitions, but not for the prediction of character-related features. Subsequently, we investigated whether there were sets of neurons that drove prediction performance, which were not apparent when considering only stimulus-aligned changes in firing activity. To address this, we identified a separate set of neurons through a logistic regression model fitted to our data and retrained our base model for the main character. Restricting the decoding to a smaller subset of these neurons lead to a slight improvement in performance, indicating that the relevant decoding information is embedded within smaller subsets of neurons. Our findings suggest that the relevant information for decoding differs by semantic category, as visual transitions relied on boundary-selective single neurons, whereas character features were carried by distributed subsets of neurons.

Results

We recorded from 29 patients (17 female, ages $22 - 63$) as they watched the movie *500 Days of Summer* (83 minutes) in its entirety (Fig. 1a). Patients were bilaterally implanted with depth electrodes for seizure monitoring, and spiking activity was recorded from a total of 2286 single- and multi-neurons across the amygdala (A; 580 neurons, 25.37%), hippocampus (H; 794, 34.73%), entorhinal cortex (EC; 440, 19.25%), parahippocampal cortex (PHC; 373, 16.32%), and piriform cortex (PIC; 99, 4.33%) (Fig. 1b). We pooled the neurons across patients (distribution shown in Supp. Fig. S1) and performed subsequent analyses on the resulting population. Due to the low number of neurons relative to the complete population, the PIC was excluded from subsequent region-wise analyses.

To determine which semantic features of the movie are represented in MTL activity, we obtained frame-wise

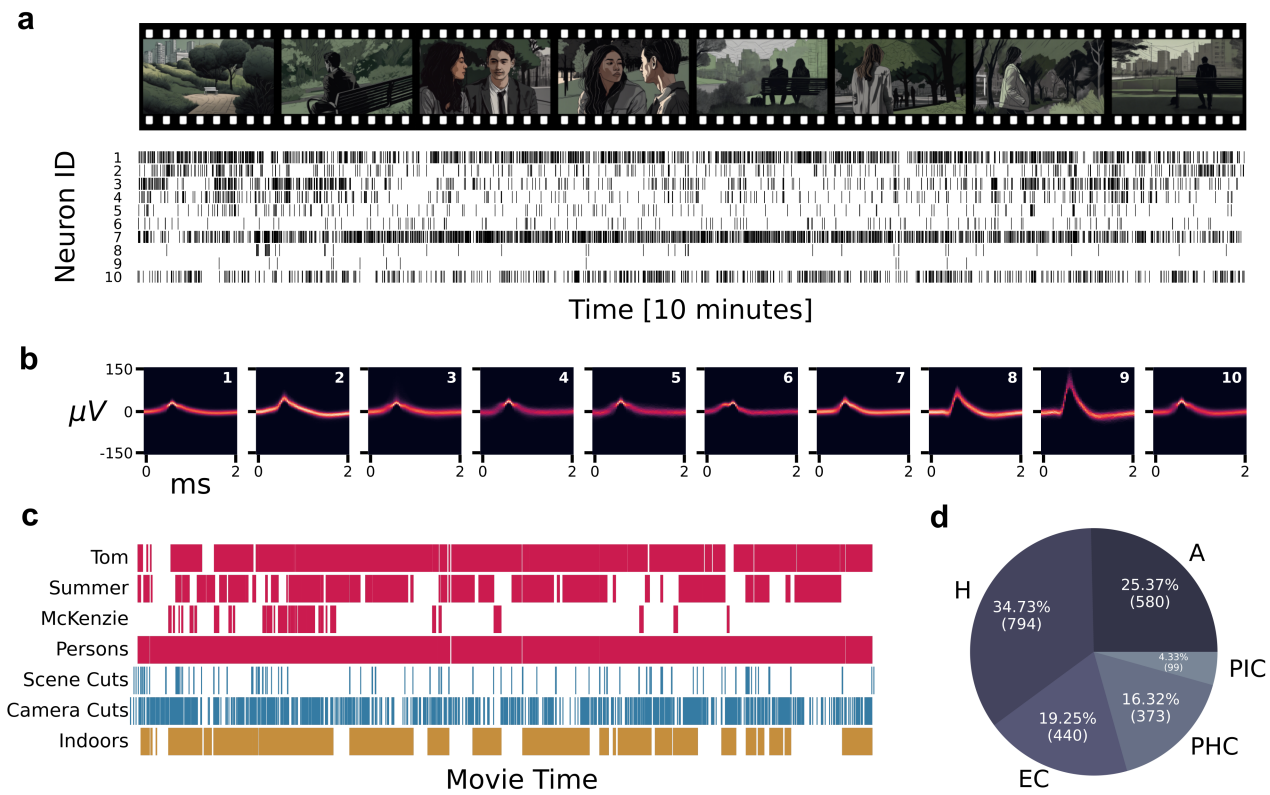

Figure 1. Overview of dataset, features and decoding approach. a) Example of the recorded neuronal activity. Patients watched the complete commercial film *500 Days of Summer* while neuronal activity was recorded via depth electrodes. Top row: example movie frames. The movie frames used in the experiment can not be displayed due to copyright and have been replaced with images generated using stable diffusion[24]. Bottom row: spike trains from 10 amygdala neurons of a single patient, where each row shows data from an individual neuron (corresponding ID number given as a label). **b)** Spike density plot showing the waveforms of each neuron in **a** (corresponding neuron ID given in top right). Neurons shown include both single- and multi-neurons. **c)** Distribution of labels across the entire movie (runtime: 83 minutes). Occurrences of character-related features are in magenta, visual transition events in blue, and location events in yellow. **d)** Distribution of the 2286 neurons across the recorded regions (A: amygdala, H: hippocampus, EC: entorhinal cortex, PHC: parahippocampal cortex, PIC: piriform cortex), for all 29 patients.

labels of several visual categories present in the movie. These categories included visual transitions, such as camera cuts and scene cuts (also referred to as soft and hard event boundaries[15]), as well as character-related features and location information (Fig. 1c). We used a mixture of manual and semi- and fully-automated methods for obtaining annotations (see Methods; sample frames in Supp. Fig. S2). For our analyses, we focused on the characters most relevant to the movie's narrative—Tom, Summer, and McKenzie.

Stimulus-aligned responsive neurons found primarily in parahippocampal cortex

To investigate whether semantic features of the movie are encoded at the single-unit level, we analyzed stimulus-evoked changes in the firing rate of neurons after the onset of characters, indoor setting, and visual transitions (Fig. 2a, Supp. Fig. S3 - S5). We classified each individual unit as responsive or non-responsive according to a previously established criterion[25], adapted to the dynamic presentation (see Methods). We found significant stimulus-evoked responses in all regions (binwise Wilcoxon signed-rank test, Simes-corrected, $\alpha = 0.001$) for visual transitions, Persons, and the characters Tom and Summer, including Face-only onsets (see Supp. Table S1). Of these stimulus-evoked changes, the parahippocampal cortex contained subpopulations of neurons which responded to the largest set of stimulus features (Summer, Persons, Camera Cuts, Scene Cuts; $p \leq 0.001$, cluster permutation test), as compared

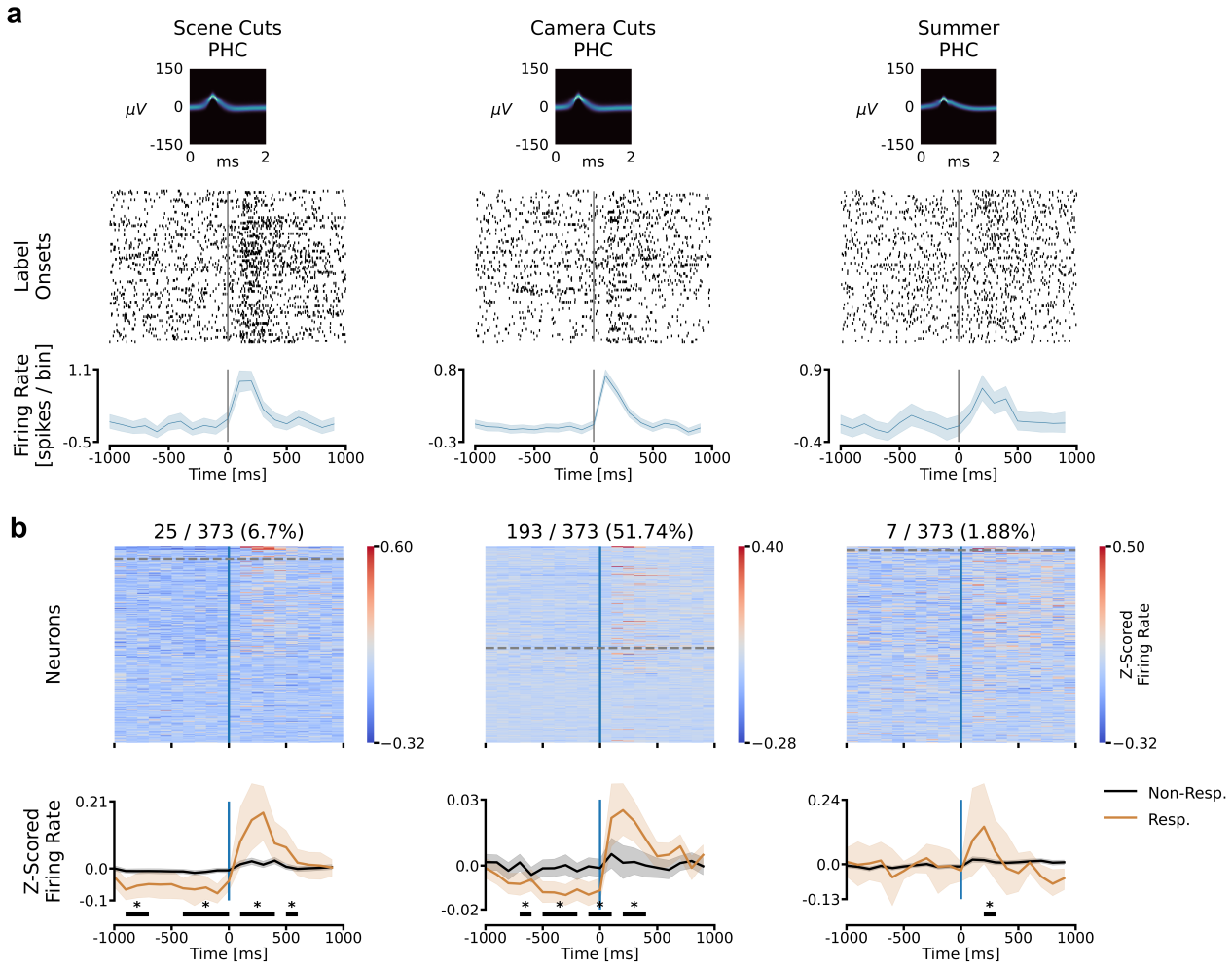

Figure 2. Responsive single-neurons in the parahippocampal cortex. a) Example peri-stimulus activity for a representative parahippocampal (PHC) neuron, for labels with a significant PHC response. Upper plots: spike density plot showing the waveforms a given neuron (label name given as title). Middle plots: spike time rasters showing the single-neuron activity surrounding the onset of the corresponding character or the visual transition event. Note: onsets for Scene Cuts and Camera Cuts were randomly subsampled to match the number of Summer appearances. Lower plots: average firing rate across 100 ms bins, for 1000 ms before and 1000 ms after the onset or event. Solid lines show the mean across all neurons, within group, and the transparent area shows the 95% confidence interval. **b)** Region-wise single-neuron activity surrounding the onset of labeled entity. Number of significantly responsive cells over the total number of PHC cells are given as the title, followed by the percentage of responsive neurons. Upper plots (heatmaps): averages of peri-stimulus spike rates per neuron (spikes per 100 ms bin, z-scored across the pseudotrial) for 1000 ms before and 1000 ms after label onset. Each row of the heatmap represents the average binned activity for one neuron. Neurons are sorted in descending order by the p-value of the response—the dotted grey line shows the threshold for responsive neurons ($p \leq 0.001$). Lower plots (line plots): average z-scored firing rate across bins. Neurons are separated into responsive (orange line) and non-responsive (black line). Solid lines show the mean for each group of neurons (responsive vs. non-responsive) and the transparent area depicts the 95% confidence interval. Significant differences between the responsive and non-responsive firing rates are shown as solid black lines (*, $p \leq 0.05$, cluster permutation test).

to the hippocampus (Tom, Camera Cuts) and amygdala (Camera Cuts). For Camera Cuts (analogous to "soft boundaries"[15]), over half of the parahippocampal neurons were responsive (193/373, 51.74%, binwise Wilcoxon signed-rank test, Simes-corrected, $\alpha = 0.001$) (Fig. 2b), and stimulus-evoked changes in firing were consistent across the set of responsive neurons ($p \leq 0.001$, cluster permutation test). For the remaining regions, stimulus-evoked modulations were less consistent across individual neurons (see Supp. Fig. S5a). For Scene Cuts ("hard boundaries"[15]), the parahippocampal cortex exhibited fewer total responsive neurons (25/373, 6.70%), but showed a similarly consistent pattern of modulation across responsive neurons ($p = 0.018$, cluster permutation test) (Fig. 2b). For characters, consistent

patterns of stimulus-evoked modulation in the responsive neurons were only observed for Summer (PHC, $p = 0.006$, cluster permutation test) and, to a lesser extent, Tom (H, $p = 0.022$, cluster permutation test) (Fig. 2a, Supp. Fig. S3). Taken together, the parahippocampal cortex contained neuronal subsets that showed a consistent pattern of increased firing after the onset of a visual transition (Camera or Scene Cut) within the movie. Characters and character faces evoked a clear change in firing activity at the subpopulation-level in the parahippocampal cortex, and sparingly in the hippocampus, but not in other tested regions.

Decoding of semantic content from population responses

The observed pattern of responses in individual neurons aligned to feature onsets suggests that these cells primarily carry information relating to visual transitions, and, to a lesser degree, the characters Summer and Tom, and Persons generally. However, previous work suggests differences in the coding capacity between the single-neuron and population level[26], and that such differences may only be apparent when analyzing specific neurons of information, such as the concurrence of active spiking periods in a subset of cells with periods of silence in others[27]. Recent theoretical work suggests that single neurons within a network can obtain high-level prediction performance for fine-grained image features, while broader categories rely on the pooling of information across neurons[28]. To explore stimulus-related information within the neuronal population, we decoded character presence, location, and visual transitions from the aggregated neuronal activity of the entire population via a recurrent neural network sensitive to both within- and between-neuronal activity patterns.

Decoding from neuronal activity

We aligned the activity of neurons across patients using the movie frames as a common time reference, generating a single neuronal pseudo-population, as done in previous work decoding from populations of single neurons[10,29]. We first assessed single-neuron decoding performance for the main character, Summer, by fine-tuning a firing activity threshold for each neuron (see Methods). This approach established a performance baseline for decoding character presence solely from single-neuron activity, without machine-learning based decoding algorithms. Decoding performances which simply apply a threshold to single-neuron activity did not reliably predict the labelled content, as most were near chance level for the majority of neurons in the population (Fig. 3b). Subsequently, we extended this analysis to a population-based approach, employing a Long Short-Term Memory network[30] (LSTM), a deep neural network well-suited for processing dynamic sequential data (Fig. 3a). This population-based approach improved decoding performance for Summer, surpassing the performances achieved by single neurons alone. We found that decoding performances exceeded chance level for all tested labels (Fig. 3c). Decoding results were consistent across metrics (F1 Score, PR-AUC, and AUROC shown in Supp. Table S2). We determined statistical significance of our results by permuting test set labels (N=1000), and found that all decoding performances were significantly above chance level at an alpha level of 0.001 (see Methods). To ensure that the significance was not due to the permutation procedure, we tested the results by using two different approaches: by shuffling randomly to destroy temporal structure and by circular shifting to maintain temporal structure, both of which produced similar results (Supp. Fig. S11). The Persons label achieved the highest performance (mean $\pm$ standard error of the mean, SEM: 0.36 ± 0.05), followed by the location-related label Indoor (0.31 ± 0.06). Character-specific labels (Summer, Tom, McKenzie) showed comparable performances, ranging between 0.23 and 0.32. Labels related to visual transitions in the movie exhibited the lowest but nonetheless significant performances of 0.2 ± 0.03 and 0.18 ± 0.01, respectively.

To test the hypothesis that response patterns are based on population activity rather than individual neurons, we also trained an LSTM network on individual neurons. From the 46 neurons identified as responsive to Summer in the separate single neuron analysis, we selected a subset of neurons, ensuring an even distribution across both patients and regions. Similar to the threshold model, most models did not reach notable prediction performances (see Fig. S6). In summary, our decoding network achieved consistent and statistically significant decoding performance on the population level exceeding chance level for all labels.

Choice of architecture.

Our pipeline uses a recurrent neural network (LSTM) to process spiking data as a time series of event counts. We binned spike counts into 80 ms intervals, covering 800 ms before and after label onset, creating sequences with a length of 20 and a dimensionality of 2286 neurons (see Methods). We compare the LSTM's decoding performance to a

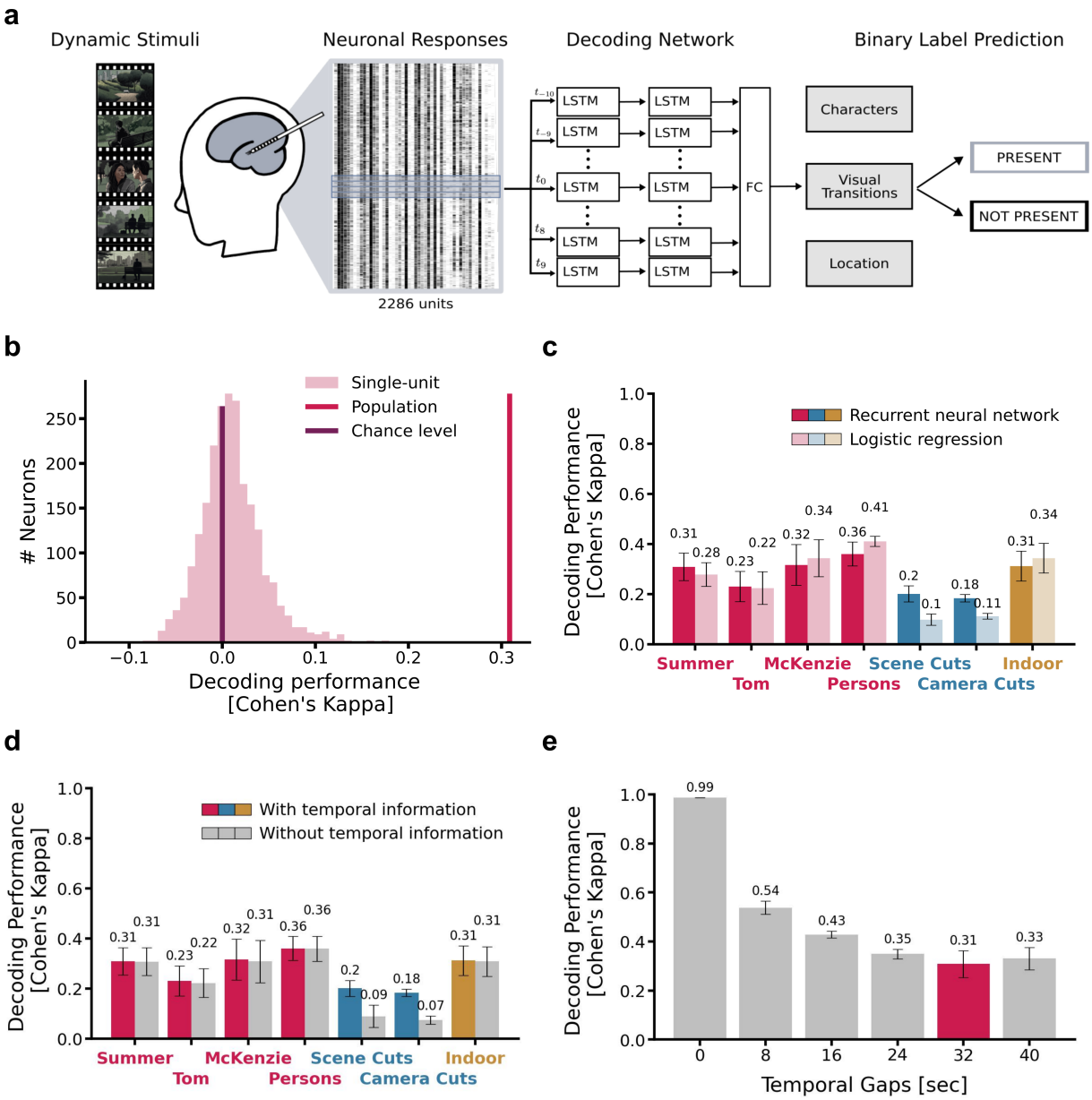

Figure 3. Categories of labels can be decoded from the neuronal population activity. a) Overview of the neural decoding pipeline. Spiking data (individual neurons shown as columns) was sectioned into 1600 ms sequences (800 ms before and after a frame onset) and given as input to a two-layer Long Short-Term Memory (LSTM) network. The output of the fully connected layer (FC) predicts the presence of a given label in a frame. **b)** Assessment of individual-neuron decoding performance by classifying data samples into positive or negative predictions for the label Summer based on the firing activity of a neuron. **c)** Decoding performances on labels of the movie (reported performance using Cohen's Kappa, mean performance across five different data splits, error bars indicate standard error of the mean). Labels fall into one of three categories — characters (pink), visual transitions (blue), or location (yellow) — with a separate model for each label. All performances were significantly better than chance level with an alpha level of 0.001. Decoding performances for the logistic regression model in lighter colors. **d)** Impact of temporal information in spike trains for recurrent neural networks. Trained models were evaluated using temporally altered test data (sequence order shuffled, repeated 100 times). Colored bars depict performance without shuffling, while grey bars represent shuffled scenarios (reported performance using Cohen's Kappa, results show the mean performance across five different data splits, variance given as standard error of the mean). **e)** Decoding performance of the main movie character (Summer) for different temporal gap sizes between samples of train, validation, and test sets. Colored temporal gap of 32 s indicates the chosen gap size for all reported performances in this study.

simpler logistic regression model, which is a linear method that does not consider the neuronal activity as a sequence of spike counts, and thus ignores the temporal information and non-linear dynamics (Fig. 3c). Apart from this, the setup and data split for both pipelines were identical. The logistic regression model showed lower performances for the Scene and Camera Cut features (by 0.1 and 0.07, respectively), whereas no drop for character-related or location-related features was observed. To test our hypothesis that the temporal information within spike sequences influences visual transition decoding, we assessed trained models using temporally-modified test data (sequence order of the spike trains was randomly shuffled, see Methods for more details) (Fig. 3d). A consistent pattern emerged, with a noticeable decline in performance, especially for visual transition labels.

Avoiding spurious decoding performance by introducing temporal gaps.

Since each frame of the movie shares a high degree of similarity with neighboring frames, we controlled for the temporal correlations in the semantic content induced by the continuous nature of the stimulus. We divided the dataset into training, validation, and test sets, with 32 s gaps between each set to minimize temporal correlations (split visualized in Supp. Fig. S7). We investigated the impact of these gaps by decoding the character Summer using varying gap lengths while keeping the number of samples comparable. We observed that smaller gaps result in substantially higher decoding performance on the test set, raising concerns about potential data leakage between train and test sets. For instance, a random split without any temporal gaps approaches a perfect score, obtaining 0.99 ± 0.0005. However, as temporal gap sizes increase to 32 s, the performance drops precipitously to 0.31 ± 0.06 (Fig. 3e; additional metrics in Supp. Fig. S8). This might explain the higher decoding performance for a comparable task in Zhang et al. [19], which did not report the use of temporal gaps for model evaluation. All subsequently reported results refer to the performance on the held-out test data using 5-fold cross-validation, with data splits incorporating the most conservative temporal gap of 32 s (see Methods). Our analysis underscores the importance of appropriate architecture selection and careful data preparation in complex datasets such as ours, as these choices can exert a significant impact on the results.

Amygdala drives decoding of character presence

The MTL has been shown to carry information about character identity, established primarily through the tuning of individual neurons [6,31,32]. To investigate character-driven activity at the population level, we analyzed neuronal activity during the presence of the movie's three main characters, Summer, Tom, and McKenzie, as well as the more general concept of any character appearance (Persons, see example frames in Supp. Fig. S2). While Summer's appearance is balanced (50/50), the remaining labels are highly unbalanced: Tom and Persons appear in the majority of frames (80/20 and 95/5), while McKenzie is predominantly absent (10/90) (Fig. 1c). Despite the imbalances, we observed significant decoding performances for all four character labels ranging between 0.23 and 0.36 (Fig. 3c).

Distribution of information across MTL regions.

To investigate region-wise differences, we trained separate decoders for neurons in the amygdala (A), hippocampus (H), entorhinal cortex (EC), and parahippocampal cortex (PHC) of the MTL. We excluded the piriform cortex (PIC) due to its low number of neurons (Fig. 1b). Our results show that all four tested regions carry information about the character's identities and reach decoding performances above chance level ($p < 0.001$, permutation test, see Methods). The amygdala and parahippocampal cortex showed the highest decoding performances for Summer and Tom, respectively, approaching levels similar to decoding from the full population. However, the distribution of information among the other regions was less consistent and varied across labels (Fig. 4a). The hippocampus had the lowest performance for Summer (0.09 ± 0.04) and Tom (0.05 ± 0.05), while the entorhinal cortex performed lowest for McKenzie (0.12 ± 0.02). For Persons, the parahippocampal cortex dominated with decoding performance comparable to the full population (0.36 ± 0.05), while other regions ranged between 0.14 and 0.21.

Differences in character's visual appearances.

Since the labels indicate a concept's presence within a natural scene rather than a single, isolated concept, the visual appearance of the labelled entity varied substantially. To analyze the differences in decoding for various levels of character appearance, we created the additional labels Summer Faces, Tom Faces, and Summer Presence (see Methods). As with the character labels, we trained a decoding network on both the full neuronal population as well as the four individual regions (Fig. 4b). Both face labels demonstrated a slight improvement on the full neuronal populations,

with Summer increasing from 0.31 ± 0.06 to 0.33 ± 0.06 and Tom increasing from 0.23 ± 0.06 to 0.27 ± 0.07. Regionally, the distribution remained consistent with the general character labels, except for Tom Faces, where the entorhinal cortex had the highest performance. The hippocampus performed weakest for both Summer Faces (0.08 ± 0.04) and Tom Faces (0.11 ± 0.03). The Summer Presence label had an overall performance of 0.27 ± 0.04, slightly lower than the general Summer label, with the amygdala clearly dominating the regional distribution (0.26 ± 0.07). For face labels, distributions were similar across regions than for the general character labels. The amygdala outperformed other regions in decoding the abstract presence of Summer but performed poorly for general person appearances, where the parahippocampal cortex performed best. These findings align with previous research showing that semantically-tuned cells in the human MTL can flexibly activate when their preferred conceptual category is indirectly invoked[33].

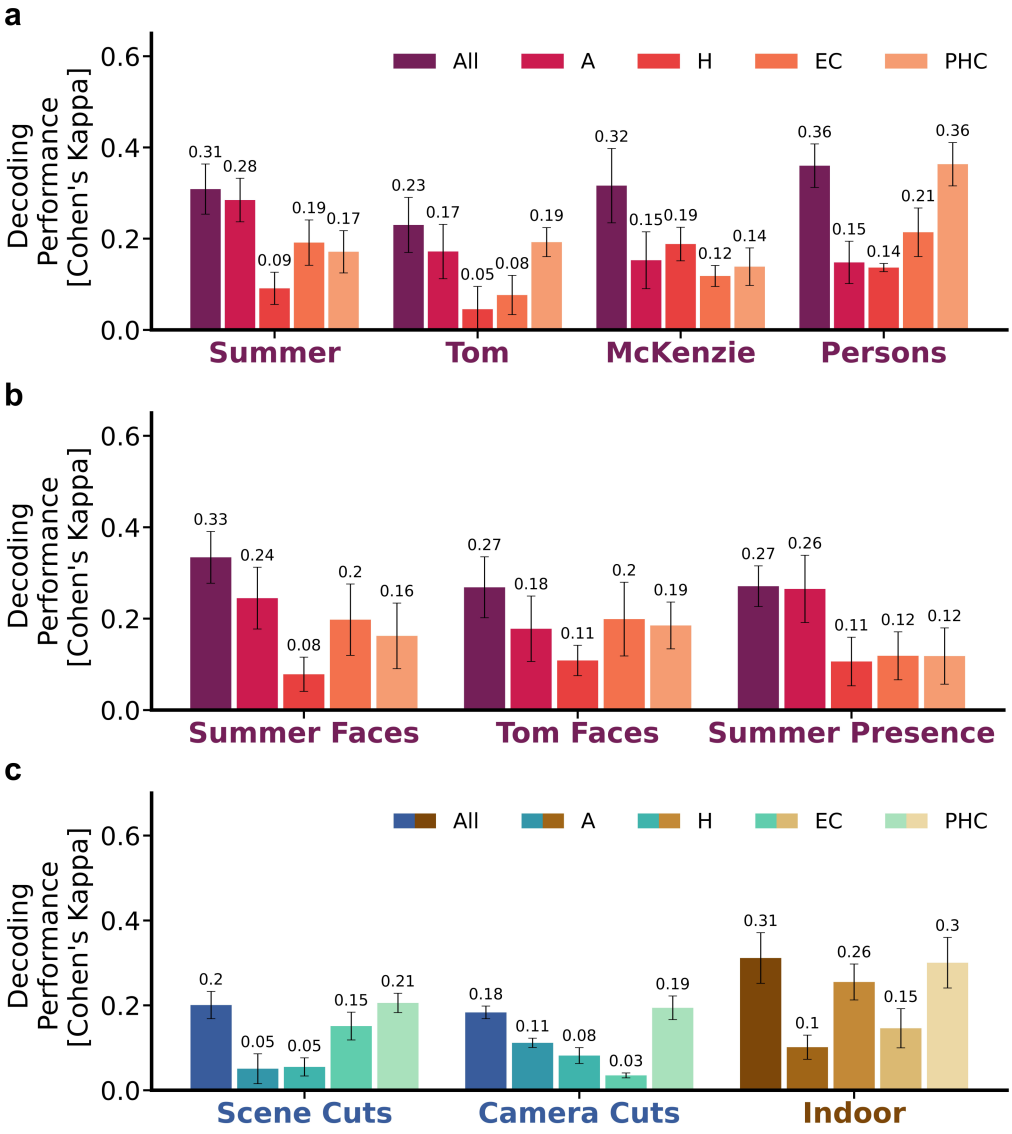

Figure 4. Semantic information is distributed differently across MTL regions based on category. Decoding performances for semantic features, by MTL region. All performances were significantly better than chance level with an alpha level of 0.001 (reported performance using Cohen's Kappa, mean performance across five different data splits, error bars indicate standard error of the mean. **a)** Character visibility could be decoded from the entire population of neurons, and with variable performance when training only on individual MTL regions. **b)** Decoding performances for face-specific character appearances and Presence features, by region. **c)** Decoding performances for visual transitions and location features.

Parahippocampal cortex drives decoding of visual transitions and location

An important visual feature in a movie is the change in camera view or scenery, which generally consists of a specific, momentary change in the visual input. Previous studies investigated comparable events consisting of rapid visual transitions in dynamic stimuli and the related modulations of neuronal activity [15,16]. We labeled the movie for changes in the visual stimuli and differentiated between a change in scenery (Scene Cuts) and a sudden change in the filming angle (Camera Cuts). Examples of Scene versus Camera Cuts are shown in Fig. S2.

To investigate whether cuts were primarily processed in a specific MTL region, we conducted a similar analysis as before by retraining on region-specific activity (Fig. 4c). We observed a clear dominance of the parahippocampal cortex for both types of visual transitions. The decoding performances reached 0.21 ± 0.02 for Scene Cuts and 0.19 ± 0.03 for Camera Cuts, respectively, when restricting the decoding to the parahippocampal cortex as opposed to 0.20 ± 0.03 and 0.18 ± 0.01 when decoding from the full population. The other regions showed a lower but above-chance decoding performance. Similarly, the parahippocampal cortex yielded the highest performance for the indoor concept and reached a performance of 0.30 ± 0.06, comparable to the performance on the full population (0.31 ± 0.06). Hippocampus was the second strongest region with a high performance of 0.26 ± 0.04. For entorhinal cortex and amygdala, we observed lower performances of 0.15 ± 0.05 and 0.1 ± 0.03, respectively.

In summary, the parahippocampal cortex consistently showed the highest decoding performance among the MTL regions, aligning with previous research in the MTL [15]. Given that both Scene and Camera Cuts are linked to sharp visual transitions, we anticipated that earlier processing stages in the MTL would show better decoding performances than later processing stages. However, despite the clear dominance of the parahippocampal cortex, our results show that other regions achieve lower, but nonetheless significant performances.

Responsive neurons drive decoding of visual transitions but not decoding of characters.

While only a subset of neurons modulated firing in response to the onset of a given movie feature, we nonetheless observed significant decoding performance from the full population. To investigate the neuronal drivers of the full population decoding performance, we defined subsets of neurons based on the responses of individual neurons and re-trained an LSTM on each subset and compared the prediction performance to that of the full population (Fig. 5).

First, we tested prediction performance using only the non-responsive neurons to determine if similar decoding could be obtained without neurons which significantly modulated firing post-stimulus onset. For this subpopulation (*Non-responsive (only)*), a separate LSTM was retrained and tested. Despite the exclusion of responsive neurons, these subpopulations yielded performances comparable to those of the complete population (Fig. 5a, *Complete*) for the character Summer. Minimal differences in performance were observed across MTL regions, with only the entorhinal and parahippocampal cortex showing qualitative decreases in performance. For comparison, we additionally retrained using only the responsive neurons as input, and again tested the decoding performance (*Responsive (only)*). In contrast to the minimal differences observed in the *Non-responsive (only)* model, restricting to only responsive neurons produced a decrease in overall performance across the entire MTL and all regions, suggesting that these responsive neurons do not drive the decoding performance observed in the full population model.

A different pattern emerged for the Camera Cuts feature, which exhibited the greatest drop in performance when excluding responsive neurons, for the entire MTL, as well as all subregions but the entorhinal cortex (Fig. 5b). This pattern was most pronounced for the parahippocampal cortex, indicating that responsive neurons carry valuable information for processing Camera Cuts in the movie. This hypothesis is further supported by the decoding performances obtained when restricting the decoding to only responsive neurons; despite the restriction in the number of input neurons, the performance only slightly dropped, and remained comparable to that of the complete population for the entire MTL, as well as for the amygdala, hippocampus, and entorhinal cortex regions.

Since the total number of responsive neurons was lower than that of non-responsive neurons, we tested the performance using size-matched versions of both non-responsive and responsive subpopulations. To match the size of the responsive neurons, we randomly selected an equivalent number of non-responsive neurons (*Non-responsive (size-matched)*) and trained and tested a separate LSTM. This process was repeated three times with different random selections, and we report the average performance. For both labels, Summer and Camera Cuts, the smaller size-matched non-responsive subpopulation showed an expected decrease in performance compared to the full non-responsive subpopulation. However, the results diverged when comparing the size-matched populations: For Summer, the size-matched non-responsive neurons performed comparably to the responsive neurons within individ-

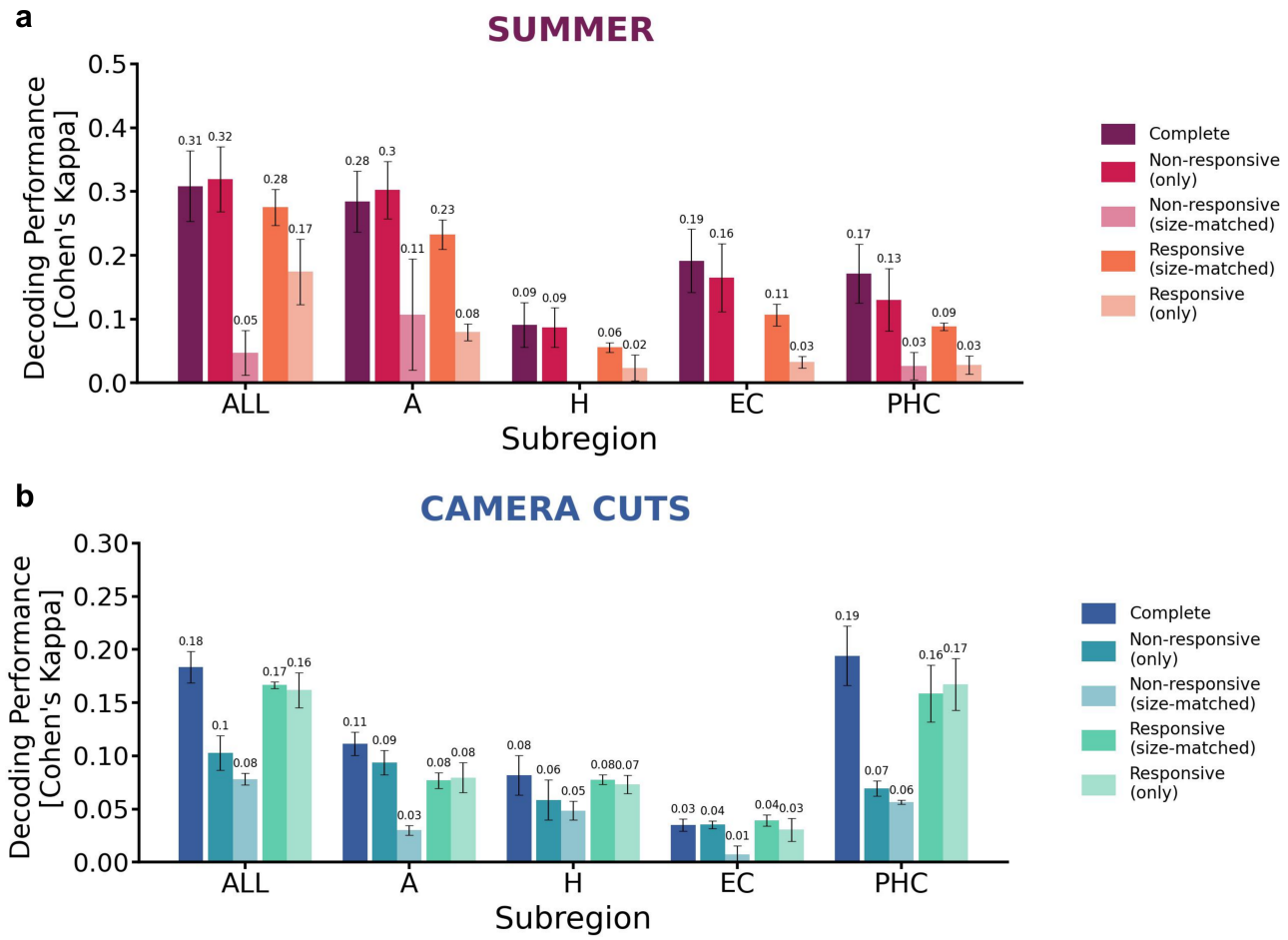

Figure 5. Responsive neurons drive performance for visual transitions, but not characters. To assess the influence of the responsive neurons (identified in Fig. 2) on decoding, we compared the decoding performance for subpopulations which did or did not contain these neurons. For the complete MTL and each individual region, the following five sets of neuronal populations were created and compared: the original complete population (*Complete*), only non-responsive neurons (*Non-responsive (only)*), randomly selected non-responsive neurons, sampled to match the number of responsive neurons, (*Non-responsive (size-matched)*). responsive neurons padded with randomly selected non-responsive neurons to match the number of non-responsive neurons (*Responsive (size-matched)*), and only the responsive neurons (*Responsive (only)*). Subpopulations with random selections were sampled three times independently, and their performance was averaged (reported performance using Cohen's Kappa, mean performance across five different data splits, error bars indicate standard error of the mean). **a)** Decoding performances for the character label Summer. **b)** Decoding performances for the visual transition label Camera Cuts.

ual MTL regions. Only for the complete population did the responsive neurons show a clear improvement. Conversely, for the Camera Cuts, restricting to only the responsive neurons improved performances, with decoding predictions of all tests surpassing those of the size-matched non-responsive set of neurons.

A similar pattern emerged for a size-matched version of the responsive neurons (*Responsive (size-matched)*), which we formed within region by padding the set of responsive neurons with randomly selected non-responsive neurons. For Summer, decoding from this subpopulation consistently showed a performance drop compared to the total set of non-responsive neurons across all tested regions. In contrast, for Camera Cuts, the size-matched subpopulation of responsive neurons achieved similar or better performance than the non-responsive neurons across all regions, with the complete population and parahippocampal region showing a clear dominance of the subpopulation containing responsive neurons. We additionally evaluated performances for the labels Tom and Scene Cuts, which showed a pattern consistent, but less pronounced, as described above (Supp. Fig. S9).

In summary, our findings indicate that responsive neurons play distinct roles for different label categories. Neu-

rons responsive to visual transitions appeared to carry information not equally present in other neurons. On the contrary, for character-related labels, individual responsive neurons contributed less to decoding, and information appeared to be distributed either across the entire population or a subset of neurons distinct from the previously identified responsive neurons.

Relevant information is carried by a smaller subpopulation of 500 neurons

Previous work in sensory information processing has found that the relevant information for a given stimulus may be encoded by only a subset of cells in a population [26,34,35]. In our above analysis, we defined subsets of neurons based on stimulus-response profiles and found, particularly for character-related labels, that such responsive neurons do not drive the decoding performance achieved by the full population. We next wanted to investigate how many neurons would be sufficient for successful decoding, and identified a more data-driven sub-selection identified by ranking the weights of a trained logistic regression model.

Training on pre-selections of top-ranked neurons.

Compared to an LSTM, logistic regression models achieved comparable decoding results for nearly all labels, except for Scene Cuts and Camera Cuts where they still showed notable decoding performance (Fig. 3c). The above LSTM and logistic regression models were trained on population data using 5-fold cross-validation, with alternating test sets for each of the five splits. However, this approach precludes an independent ranking of neurons across splits, as test data from one split may overlap with training data from another. To address this, we expanded the five original splits to 20, where each subgroup of four splits shared a common test set but varied in the allocation of training and validation data. Any analysis relying on a subset of neurons derived from logistic regression weights was exclusively assessed using the four splits that produced the ranking and shared held-out test data (see more details in Methods, and Supp. Fig. S10).

We trained on progressively smaller subpopulations of top-ranked neurons for the character Summer (Fig. 6a) and observed an increase in performance when restricting the input activity to smaller populations (peak at 0.38 (Cohen's Kappa) for 500 neurons, 21.9% of the total population). The decoding performance reported here for the entire population shows a slight variation from the previously reported value of 0.31 due to the modified nested cross-validation procedure. Further reduction of the population lead to a decrease in performance, yet high decoding performance persisted even in small subpopulations of neurons.

We observed that a ranking procedure which did not use a common test set was subject to potential cross-talk between datasets, which significantly impacted and distorted the results. Using a selection of neurons derived from non-independent train and test data led to a stark increase in performance, nearly doubling the original performance of the character Summer to 0.52 (Cohen's Kappa) when restricting to a subpopulation of 150 top-ranked neurons. This again underscores the need to carefully prepare the data for paradigms like ours, where data samples are highly correlated, as ignoring dependencies between train and test data can greatly affect the results (see Fig. S13).

Top-ranked neurons and their distribution across patients and MTL regions.

As the selection process involved five distinct rankings of neurons, we considered how consistent the rankings were. We illustrated the overlap of neurons within subpopulations of top-ranked neurons assessed across different sizes (Fig. 6b). Analyzing the top-performing 500 neurons from each of the five rankings revealed a common set of 78 neurons. We compared this observed overlap to that expected by chance with random subpopulation selections (see Methods), finding a notably higher overlap. This suggests the presence of common neurons crucial for decoding the label Summer. Additionally, we also compared these selected neurons to those previously classified as responsive. The overlap between the two groups of neurons was small, suggesting that the responsive neurons defined on a basis onset-related changes in activity may not be the primary drivers of decoding performance in general, aligning with previous observations for character labels when the decoding pipeline was restricted to responsive neurons. As restricting to 500 neurons yielded the highest decoding performance, we subsequently analyzed the resulting intersection of 78 neurons across rankings. These 78 neurons were distributed across both patients and regions (Fig. 6c,d), with no single patient or area of the MTL driving the performance. Visualizations of the spiking activity surrounding the onset of Summer for these do not reveal a clear pattern of stimulus-evoked increases in firing (Supp. Fig. S12).

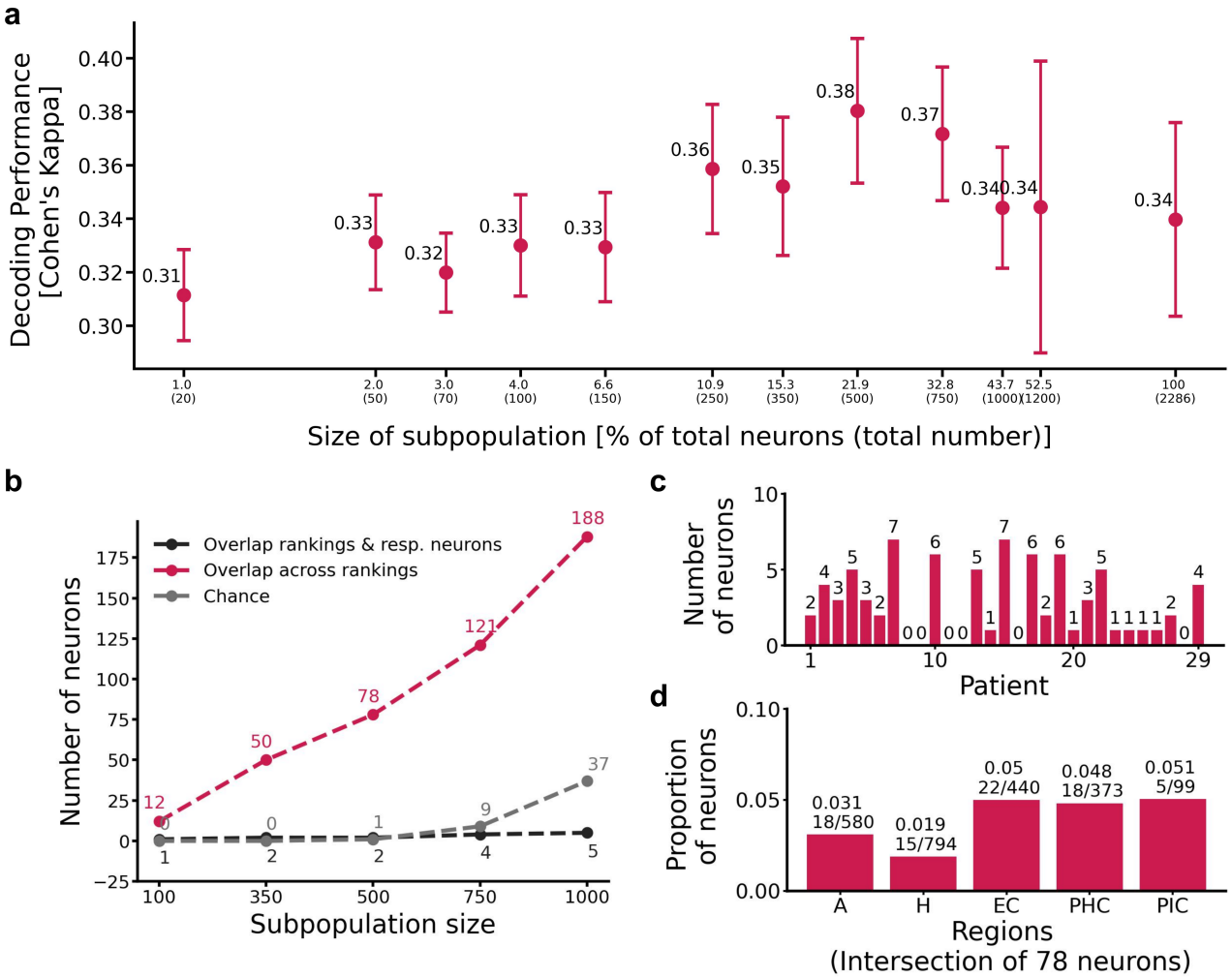

Figure 6. 500 neurons are sufficient to reach peak decoding performance. a) Decoding performances for the character Summer for subpopulations of top-performing neurons, testing various sizes ranging from 1% to 100% of the full population (absolute numbers of neurons are reported in parenthesis). Mean performance across different splits is reported, and standard error of the mean is visualized by the error bars. **b)** Number of overlapping neurons across rankings for different sizes of subpopulations of top-performing neurons (pink). As a baseline, we compare the number of overlapping neurons to the number expected by chance (grey), and we observed a notably higher intersection of top-ranked neurons across the splits. Additionally, the overlap between the intersection of top-ranked neurons and the previously defined responsive neurons is shown (black). **c, d)** Overlapping neurons (in total 78) in subpopulations of 500 top-performing neurons for each ranking were distributed across patients and MTL regions.

Our results support the idea that the full population contains redundant and noisy information which is unnecessary for the decoding task. We identified the relevant subpopulation as a subset of approximately 500 neurons and showed that these neurons are distributed across patients and MTL regions.

Discussion

We investigated how the brain processes semantic information in a naturalistic setting by analyzing the activity of single neurons in the human MTL during the presentation of a full-length movie. Our contributions are three-fold: First, we showed that the information necessary to decode stimulus features depends on the semantic category, as the decoding of characters and location relied on the population while the prediction of the visual transitions was carried by visually-selective individual neurons. Second, we found that a given concept was represented by a smaller subpopulation of the total population of recorded neurons not apparent from the single-neuron activity alone, and that the distribution of information across neurons differed between MTL regions based on the type of

stimulus. Finally, our results extend the region-specific findings from the static to the dynamic regime by showing that character decoding was primarily carried by the amygdala[6], whereas transitions between camera angles and scenes were largely carried by the parahippocampal cortex[36].

By leveraging different decoding architectures, we were able to further investigate the temporal dynamics underlying the representation of semantic features. Through a comparison between decoding architectures, we found that temporal information in the spike trains only improved prediction of visual transitions, while ignoring sequence information had little effect on character and location features. Though utilizing temporal dynamics enhanced decoding performance for visual transitions, the temporal dynamics in our data, particularly those inherent to the movie stimulus, also present a significant confounding factor. Both the movie and recorded neuronal activity are subject to a high degree of correlation in time and thus required extra consideration when setting up the decoding network to ensure that the train and test data did not contain adjacent, highly-correlated samples. Correlation between train and test samples artificially inflates decoding performance, and does not reflect actual generalization to unseen data. In related research by Zhang et al.[19], where temporal distance was not accounted for, high decoding performances were reported on a similar task. Based on our analyses (Fig. 3e), we anticipate that their reported decoding accuracy is overestimated, and would change if sufficiently large gaps were introduced.

Additionally, our findings suggest that the informative unit of neural activity may differ based on stimulus category. We found that for visual transitions, the decoding performance was driven by a subpopulation of responsive individual neurons, whereas for character features, decoding performance worsens when restricted to individually responsive neurons. For the visual transitions, this reliance on a subset of responsive neurons is most pronounced in the parahippocampal cortex (PHC), as performance dropped considerably upon the exclusion of responsive cells and was on par with a random selection of non-responsive neurons. This effect may reflect a general scene-selective preference of PHC neurons, as opposed to a more invariant and specific tuning, supporting previous work on the same subject[36]. Notably, even with this apparent reliance on feature-selective cells, the overall prediction performance was lower for visual transitions generally, as compared to that of the characters.

Using the neuronal ranking obtained from logistic regression models, we pursued a more data-driven approach to identify informative subsets of neurons. Compared to decoding from the full population, which may have introduced noisy information, we found an increase in decoding performance when restricting to a subpopulation of top-performing neurons. More importantly, for these specific neurons we found that there was not a consistent pattern in the response elicited by the stimulus onsets within the movie. These results suggest that the dynamically changing semantic information is not totally dispersed throughout the population activity, but relies on a smaller subpopulation within the entire group of neurons, and that this subset does not overlap with cells showing post-onset increases in firing. It remains an open question as to whether the information decoded here strongly contributes to conscious perception.

As our dataset consists of single-neuron activity pooled across 29 subjects, each with activity from approximately 100 recorded neurons, the extent to which claims can be made about what a given, individual brain does are limited. Additionally, the patients watched a full-length movie in an clinical setting, where neuronal activity is not solely focused on visual stimuli and likely processes additional information. While we cannot know whether the neuronal population that we sampled is representative of the human MTL generally, it is one of the largest samples collected to date, both in terms of neurons and patients and offers a unique opportunity to understand the processing and representation of information in the MTL population activity. Despite these limitations, which make near-perfect performances in the decoding task unrealistic, the presented decoding approach demonstrates that movie content can nonetheless be successfully decoded from such a sub-sampled population of neurons. This successful decoding underscores the breadth of information contained within the MTL, highlighting its crucial role in processing complex visual stimuli.

Materials and Methods

Participants and recording

The study was approved by the Medical Institutional Review Board of the University of Bonn (accession number 095/10 for single-unit recordings in humans in general and 243/11 for the current paradigm) and adhered to the guidelines of the Declaration of Helsinki. Each patient gave informed written consent for the implantation of micro-wires and for participation in the experiment.

We recorded from 46 patients with pharmacologically intractable epilepsy (ages 19 - 62, median age 37; 26 female patients). Patients were implanted with depth electrodes[31] for locating the seizure onset zone for later resection. Micro-wire electrodes (AdTech, Racine, WI) were implanted inside the shaft of each depth electrode. Signals from the micro-wires was amplified using a Neuralynx ATLAS system (Bozeman, MT), filtered between 0.1 Hz and 9000 Hz, and sampled at a rate of 32 kHz. Spike sorting was performed via Combinato[12] using the default parameters and removal of recording artifacts such as duplicated spikes and signal interference was done using the Duplicate Event Removal package[37]. After all data were processed, neuronal signals, experimental variables, and movie annotations were uploaded to a tailored version of Epiphyte[38] for analysis. Due to disruptions in the movie playback caused by clinical interruptions, 13 patient sessions were excluded from further analysis.

Task and stimuli

Patients were shown a German dubbing of the commercial movie *500 Days of Summer* (2009) in its entirety (83 min). This film was chosen because the main actors were relatively unfamiliar to a general German audience at the time of the initial recordings.

The movie was shown in an uncontrolled clinical setting, where neither gaze nor attention were directly monitored, and participants could pause the movie at will. The movie was presented in letterbox format without subtitles on a laptop using a modified version of the open-source Linux package, *FFmpeg*[39], with a frame rate of 25 frames per second. Due to the length of the movie and the possibility of clinical interruptions, patients were allowed to freely pause the playback. Discontinuity in playback was controlled for within Epiphyte[38]. Pauses and skips in the movie playback were identified through the output of the modified *FFmpeg* program and used as a basis of exclusion for patients. A patient was excluded if they did not watch the entire movie, or watched the movie discontinuously.

Before and after the movie, patients completed a short screening consisting of 42 images: 32 images showing characters from the movie and 10 images from unrelated well-known films (e.g., *Titanic*, *The Godfather*). For the analysis presented in this paper, only the information for the movie portion of the experiment was used.

Movie annotations

In order to relate the content of the movie to the recorded neuronal activity, we labeled various features on a frame-by-frame basis. These labels are binary and cover the following features:

- Main characters: Summer, Tom, McKenzie
- Faces: Summer, Tom
  These labels focus on face appearances. Positive samples are frames where the character's face is shown, while negative samples are frames where the character's face is not visible at all. All other frames are excluded.
- Presence: Summer
  The presence label indicates the general presence of the character in the scene, even if not visible on the frame.
- Scene Information: Camera Cuts, Scene Cuts
  These labels mark visual transitions in the movie. Whereas Scene Cuts rely on a change of scenery, Camera Cuts are primarily based on changes in the visual stimuli.
- Location: Inside/Outside
- Persons: General appearance of any person(s)

Main character, Presence, Persons, Location, and Scene Cut labels were obtained manually using the open-source annotation program *Advene*[40]. For characters, we labeled a frame as containing that character if the annotator could clearly distinguish the actor by either appearance or context. Characters and persons were considered only on a visual basis (i.e., a frame in which Tom is speaking but not visible is labeled as not containing Tom). For face labels of the characters, we developed a deep-learning pipeline for face detection and classification. As backbone we used a

pre-trained neural network for face detection and feature extraction[41]. We extended the pipeline by a classification network consisting of fully-connected layers combined with ReLU activation functions. The classification network was fine-tuned on the movie frames to classify detected faces of the main characters, including a "not known" class for faces not belonging to the main characters. In the fine-tuning process, we used the manually created character labels for the characters. Camera Cuts were labeled automatically using the open-source algorithm *PySceneDetect*[42], run with default parameters and manually reviewed. Camera Cuts mark a cut in the movie by labeling the first frame after cut onset as positive, resulting in cut events associated with a single frame. To adjust for the temporal latency in brain activity, cuts in the movie were associated with frames occurring within 520 ms of the cut onset. This adjustment smooths the cut labels, rendering them more comparable to what we anticipate in neuronal responses.

Calculation of single-neuron response statistics

The "baseline" period was defined as 1000 ms prior to the onset (e.g., the entry of a character into frame) and the "stimulus" period was 1000 ms after the onset or appearance. Pseudo-trials with baseline periods containing frames depicting the label were excluded. Responsive neurons were defined using a modified bin-wise signed-rank test[25]. The spiking activity across pseudo-trials was aligned to the stimulus onset. The baseline period was binned by 100 ms, and the normalized firing rate of the baseline period was compared to nineteen overlapping 100 ms bins defined across the stimulus period using a Wilcoxon signed-rank test (alpha = 0.01, SciPy wilcoxon[43], Simes corrected[44]). Additionally, a neuron must spike during at least one third of the total pseudo-trials to be tested (otherwise, assigned $p = 1$).

Cluster permutation test

A cluster permutation test[45] was used to test the difference in firing rates between the responsive and non-responsive subsets of neurons. Using the calculated response statistics, neurons were divided into two conditions: responsive ($p \leq 0.01$) and non-responsive ($p > 0.01$) for a given label. Activity for a neuron was averaged across bins, yielding a single vector of mean spike counts (spikes / 100 ms) spanning both baseline and stimulus periods for each neuron. This vector was then z-scored relative to its own mean and standard deviation. Mean spike count vectors were combined across conditions, yielding two datasets: A_{resp} and A_{non}, matrices containing the summarized, binned activity for all responsive and non-responsive neurons, respectively. A bin-wise comparison between A_{resp} and A_{non} was performed using a two-sided t-test for independent samples (`ttest_ind`, SciPy), producing a t-stat and p-value for each bin. Clusters were defined as temporally adjacent bins with $p \leq 0.005$ and t-stats were summed within clusters. The procedure was adapted to allow testing of multiple clusters, so no clusters were excluded at this stage. One set of 1000 permutations were performed by randomly assigning neurons to A_{resp} and A_{non} such that the total number of neurons in each group was conserved and the bin-wise testing procedure was performed on each permuted dataset. Each cluster was compared to the resulting histogram of permuted t-stats. P-values for each cluster were defined as the number of permutations with a higher summed t-stat relative to the total number of permutations, and a p-value less than 0.05 was considered significant.

Decoding from population responses

The movie had a frame rate of 25 frames per second, such that each frame was presented for 40 ms. In total, there were 125,743 frames, and the spiking activity of the recorded neurons was binned using a bin size of 80 ms, with one bin representing the activity of two consecutive frames. For the decoding process, only the story content of the movie was used, and the credits were excluded from the dataset. The label assigned to a bin was that of the first frame. We sampled activity of the neurons before and after the onset of each frame and found that using a window of 800 ms before and 800 ms after onset produced the best results. This resulted in a data sample with a binary label for the concept and a spike train of length 20 (10 bins before and after the onset of the frame) and dimension 2286 (since the full population of recorded neurons consists of 2286 neurons).

Architecture

In order to decode concepts from the sequence of neuronal activity, we used a long short-term memory network[30] (LSTM), which is well-suited to process the dynamical structure of the dataset. The output of the LSTM was fed into into several fully-connected layers with ReLU activation functions to obtain binary label predictions. We found it beneficial

to pre-process the raw spiking data with a linear layer of same size as the input, combined with a batch normalization layer. We used the binary cross-entropy loss function, and optimized our network using Adam optimizer[46] in Pytorch[47] with default settings (first and second order moment equal to 0.9 and 0.999, respectively). We obtained the best results with a 2-layer LSTM, with hidden size of 32. We adapted other hyperparameters, e.g., number of linear layers, hidden sizes, batch size, learning rate, weight decay, and dropout rate, for each label. We trained each network for 700 epochs and used the validation set to approximate the model's capability to generalize on unseen data. As is common practice, we selected the best-performing model on the validation set to evaluate on the test set. Some of the labels were highly imbalanced and models trained on these labels were biased towards predicting the majority class. In order to ensure unbiased predictions and maximize the performance, it is common to force the batches of data samples presented in the training process to be balanced by oversampling the minority class. We applied the oversampling technique for all unbalanced labels.

Data split

To ensure that the decoding performance was not affected by correlations between data samples, we carefully split the dataset into training, validation, and test sets. We used 5-fold nested cross-validation, and assigned 70% of the data to training, 15% to validation, and 15% to testing in each split. To avoid correlations between samples, we assigned samples from consecutive segments of the movie to each set (train/val/test) and excluded 32 seconds of data between each set (see Fig. S7). The choice of excluding 32 s was based on the results for the main character Summer (Fig. 3c). The total dataset contained $45,200$ samples, resulting in sets of $30,800/7,200/7,200$ samples for training/validation/test. We trained a network using the training data, optimized its hyperparameters using the validation set, and evaluated its performance on the corresponding test set. We report the final decoding performance as the average performance on all five test sets.

Evaluation metrics

We compared the model's prediction to the binary class labels for the respective semantic concept. While accuracy is a simple evaluation metric, it is not suitable in our case because the training data for a majority of the labels is highly imbalanced. Here, we present all decoding performances using the Cohen's Kappa metric[48] which measures the agreement between the ground-truth labels and the predictions of the network, where a performance equal to zero is interpreted as chance-level and a performance equal to one is interpreted as complete agreement. Cohen's Kappa is defined as

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

where p_0 defines the relative observed agreement and p_e the hypothetical probability of chance agreement among prediction and labels. The Cohen's Kappa metric can take negative values, too, implying that the predictions are worse than chance level. The common chance performance equal to zero makes the Cohen's Kappa a useful metric to compare performances across all labels.

Additionally, we report the F1 Score, Area under the Precision Recall Curve (PR-AUC) and Area Under the Receiver Operating Characteristic (AUROC) metric in the Supplementary for all experiments. We briefly explain these metrics, including potential advantages and disadvantages for our analysis:

F1 Score

F1 Score is a metric combining precision and recall by calculating the harmonic mean between these two. The calculation of precision and recall relies on a classification threshold of 0.5. The F1 Score has been designed to work well on imbalanced datasets, which makes it a valuable metric to use in the evaluation of our decoding tasks.

PR-AUC

The PR-AUC metric as the area under the precision-recall curve is an extension of the F1 Score to various threshold settings. Similarly to the F1 Score, it can be used for imbalanced datasets. However, for both F1 Score and PR-AUC the metric are sensitive to changes in the class distributions, and hence chance performance varies across labels. Being a reliable performance metric in general for our various classification problems, one has to bear in mind that a comparison of performances across concepts can be misleading due to the different chance baselines.

AUROC

The Receiver Operating Characteristics (ROC) curve represents the trade-off between the true positive rate and false positive rate for various threshold settings. AUROC as the area under the ROC curve is a performance measure that is used in settings where one equally cares about positive and negative classes. Performances range in $[0, 1]$ and it is insensitive to changes in the class distribution, which means chance performance is given by a value of 0.5. However, the metric is generally not used for highly imbalanced classification problems and an evaluation of specific labels in our analysis such as McKenzie (class distribution is $90/10$) should be taken with caution.

Single-neuron decoding performance

For a fair assessment of decoding from population responses compared to single-neuron activity, we evaluated the decoding performance of individual neurons under a setup comparable to that of the LSTM-based decoding network. Instead of relying on full-population responses, we based predictions solely on the firing rates of a single neuron. Employing the same data preparation procedures used for the network, we summed the binned firing rates of a neuron around the onset of a frames for the 1600 ms time frame utilized by the decoding network. We individually optimized the threshold for each neuron's activity on the validation set, which is used to tune the hyperparameters during the network's training. Subsequently, we used the fitted threshold to evaluate the neuron's performance on the hold-out test set and report the final performance as the average of the five results obtained from the 5-fold cross-validation procedure.

Permutation tests for decoding

To determine if the reported decoding performances were significantly better than chance, we performed two sets of permutation tests – first, we randomly shuffled the labels of the held-out test set (here called "Test Set Shuffle"), and second, we shifted the labels while preserving the order of the test labels (here called "Circular Shift"). For both tests, the input to the decoding network remained unchanged from the non-permuted version. The only modifications made were to the corresponding feature label in the test set, which were changed in the ways outlined above, and then compared to the original network's prediction scores. By modifying only the labels, we ensured that the temporal information embedded in the neural data remained unaffected by the permutation test. The dynamic nature of the visual stimuli implies not only a strong correlation within the neural activity but also suggests a temporal correlation for the feature labels. This led us to test the significance of both scenarios: one where the temporal correlation within the labels is disrupted (Test Set Shuffle), and another where the temporal correlation is maintained (Circular Shift).

Test set shuffle.

The first type of permutation, and the one reported in the main text, consisted of randomly shuffling labels in the test set and evaluating the predictions of the models on those. We compared the performance of the model on the held-out test set to a null distribution generated by evaluating the model on the test set with shuffled labels (N=1000). We calculated the probability of our observed performance under the null distribution to obtain the p-value:

$$p = \frac{k+1}{N+1}$$

where k is the number of performances on permuted data outperforming the ground-truth performance of the model. This p-value provided the basis of comparison for describing significant decoding results in our main analyses. By preserving the temporal structure during the model's inference process and only disrupting temporal correlations within the concept labels used for evaluating the model's predictions, we consider this assessment of significance to be the most equitable for our data setup.

Circular shift.

The second type of permutation was performed as a comparison for the above *Test Set Shuffle*, as it is a standard and frequently employed method in human single-neuron studies (see [37,49,50]). Circular shift permutations maintain both the temporal relationship of neural data as well as that of the stimulus information, here the concept labels. Rather than randomly shuffling the labels in the test set, we applied a circular shift of labels that maintained the intrinsic temporal correlation. The shifting size was randomly chosen $N = 1000$ times to obtain a null distribution akin to the previous test. This distribution was then utilized to compute the probability of our ground-truth performance and

derive the p-value for assessing significance. For studies involving static stimulus presentation, wherein the stimulus is largely uncorrelated with itself, this test provides a useful way to disentangle stimulus-related effects from those endemic to the time-series information. For a comparison between the permutation results, see Supplementary Materials, Fig. S11.

Impact of temporal information on the decoding

To assess the significance of temporal information in neuronal activity sequences, we evaluated the trained models using temporally-altered test data. The input to the decoding network included neuronal activity from 800 ms before to 800 ms after the onset of the frame, divided into 20 bins. Each data sample had a dimension of 2286, corresponding to the number of neurons in our population. We randomly shuffled the sequence order of the 20 bins (applying a consistent permutation for all neurons and test samples) and evaluated the pre-trained models on the modified test data. This procedure was repeated 100 times, and the performance was averaged. The final performance was calculated by averaging the results across the five data splits, using the standard error of the mean (SEM) as the measure of variance.

Logistic regression models and evaluation of neuron contribution to decoding

We compared the decoding performance of the LSTM to that of a logistic regression model. The dataset used to train the logistic regression model was identical, barring one key change: the spike trains provided to the LSTM, initially of dimension $(20, 2286)$, were reduced to a single bin representing the total number of spikes around a frame. The data then had a revised shape of $(1, 2286)$, and no longer incorporated temporal information. We trained a logistic regression model using the `liblinear` solver implementation in Scikit-learn[51]. For training, we employed an L1 penalty and z-scored the neuronal activity per neuron using the mean and standard deviation of the training data. We utilized a nested 5-fold cross-validation, and separately optimized the regularization strength for each data split. The final decoding performance was calculated by averaging the performance across all five test sets.

Applying an L1 penalty during training enforced feature sparsity, which facilitated the interpretation of input feature importance through the model's coefficients. A ranking of neurons was generated by evaluating the coefficients of the trained logistic regression models. Caution is needed when combining logistic regression weights from models trained on different splits due to the alternating test sets in each split, which sometimes include data used for training in another split (see Fig. S7). To avoid any interference between train and test data across splits while still accounting for the temporal variation in our data, we further modified the splits used for cross-validation. The original five splits were extended to 20 splits in the following way: each of the original five splits was further divided into four sub-splits which shared a common test set but alternated the division of training and validation data (see visualization in Fig. S10). Any analysis based on a selection of neurons from logistic regression weights was exclusively evaluated on the corresponding four splits that generated the ranking and shared held-out test data. Final decoding performances for such analyses were derived through a nested cross-validation procedure. This involved initially averaging the decoding performances of models that shared a common test set (i.e. averaging across each set of four sub-splits) and then averaging the resulting five performances. In short, our training procedure for the logistic regression analysis consisted of training $5 \times 4 = 20$ splits, and thus 20 models, where each group of 4 sub-splits shared a common test set.

Neurons were ranked according to their logistic regression coefficients across each set of four sub-splits. Given that there are a total of five such groups, this lead to five distinct rankings. To obtain each ranking, we combined the coefficients of the four trained models using a two-step procedure:

1. Partition the neurons into separate subsets, based on the number of models for which a neuron had a non-zero coefficient (e.g. one group of neurons which had non-zero coefficients on all four sub-splits, then all three sub-splits, etc.).
2. Within each subset of four models, use the average of the absolute coefficients across the five splits to obtain a subset-specific ranking.

By concatenating the partitions of ranked neurons, we obtained a comprehensive ranking of all neurons. The neuron ranked highest displayed non-zero coefficients in all four models (corresponding to four sub-splits) and possessed the greatest average absolute coefficient value among neurons activated in all four models.

Intersection of top-performing neurons and chance-level overlap

In our analysis, we restricted the decoding to subpopulations of neurons that were derived through a ranking of weights of a trained logistic regression model. To ensure that the selection of neurons was independent of the test data, the neuron selection procedure was based on five rankings derived from distinct data splits, each paired with fixed test data. We evaluated the intersection of top-performing neurons across subsets of top-ranked neurons from the five rankings, evaluated for varying subpopulation sizes (Fig. 6b). For instance, a comparison of the top-performing 500 neurons from each of the five rankings revealed a set of 78 common neurons.

As a reference point for comparison, we report the average count of overlapping neurons anticipated when randomly selecting sets of 500 neurons five times from the entire population (denoted as chance level of overlapping neurons in Fig. 6). In the previously mentioned scenario involving subpopulation sizes of 500, the expected number of overlapping neurons is equivalent to one. Mathematically, this is computed as follows: the full population consists of $N = 2286$ neurons. We refer to the size of the subpopulation as $k = 500$ and the number of total rankings $m = 5$. For each neuron n_i in the subpopulation, we define a random variable $X_{i,500}$ as follows:

$$X_{i,500} = \begin{cases} 1, & \text{if neuron } n_i \text{ lies in all five subpopulations} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

We observe that $\mathbb{P}(X_{i,500} = 1) = \left(\frac{k}{N}\right)^m = \left(\frac{500}{2286}\right)^5$. The expected value of $X_{i,500}$ is given by

$$\mathbb{E}(X_{i,500}) = 1 \cdot \mathbb{P}(X_{i,500} = 1) + 0 \cdot \mathbb{P}(X_{i,500} = 0) = \mathbb{P}(X_{i,500} = 1)$$

We define $X_{500} = \sum_{i=1}^{N} X_{i,500}$ as the number of overlapping neurons across all five rankings. Since the random variables are independently and identically distributed, for the random variable X_{500} this implies

$$\mathbb{E}(X_{500}) = \sum_{i=1}^{N} \mathbb{E}(X_{i,500}) = \sum_{i=1}^{N} \mathbb{P}(X_{i,500} = 1) = 2286 \cdot \left(\frac{500}{2286}\right)^5 \approx 1.1443$$

Analogous calculations for $k = 100, 350, 750, 1000$ result in

$$\mathbb{E}(X_{100}) = 2286 \cdot \left(\frac{100}{2286}\right)^5 \approx 0.0004$$
$$\mathbb{E}(X_{350}) = 2286 \cdot \left(\frac{350}{2286}\right)^5 \approx 0.1923$$
$$\mathbb{E}(X_{750}) = 2286 \cdot \left(\frac{750}{2286}\right)^5 \approx 8.6896$$
$$\mathbb{E}(X_{1000}) = 2286 \cdot \left(\frac{1000}{2286}\right)^5 \approx 36.6180$$

Thus, we derive the chance baselines as 0, 0, 1, 9, and 37 for subpopulation sizes of 100, 350, 500, 750, and 1000, respectively.

Decoding on regions of the MTL

We compared the decoding performance when using the activity of all 2286 recorded neurons to the performance when only using activity from specific regions of the MTL. These regions are the amygdala (580 neurons), hippocampus (794), enthorinal cortex (440), and parahippocampal cortex (373). To use activity from a particular region, we limited ourselves to the activity of neurons in that region and reduced the input dimension to match the number of neurons in the region. The network architecture and data splits remained the same as when using activity from the full population, but we tuned the hyperparameters for the reduced dataset. We kept equal proportions for the training, validation, and test sets. In other words, the difference between decoding from the full population and the regions was the reduced dimensionality of the input data, which goes from a spike train of length 20 and dimension 2286 to a spike train of the same length but a reduced dimension based on the number of neurons in the region.

Acknowledgements

We would like to thank Tamara Müller and Aleksandar Levic for their contributions to the data analysis framework. This work was supported by the German Federal Ministry of Education and Research (DeepHumanVision, FKZ: 031L0197A-C; Tübingen AI Center, FKZ: 01IS18039), as well as the German Research Foundation (DFG) through Germany's Excellence Strategy (Cluster of Excellence Machine Learning for Science, EXC-Number 2064/1, PN 390727645) and SFB1233 (PN 276693517), SFB 1089 (PN 227953431), SPP 2411 (PN: 520287829), and MO 930/4-2.

Data Availability Statement

The data will be made fully available in a separate publication following the release of this manuscript. Code will be made available on GitHub.

Author Contributions

Conceptualization: Franziska Gerken, Alana Darcher, Pedro J Gonçalves, Rachel Rapp, Ismail Elezi, Johannes Niediek, Stefanie Liebe, Jakob H Macke, Florian Mormann, Laura Leal-Taixé
Data acquisition: Alana Darcher, Johannes Niediek, Thomas P Reber, Marcel S Kehl, Stefanie Liebe, Florian Mormann
Data curation: Alana Darcher, Marcel S Kehl
Methodology: Franziska Gerken, Alana Darcher, Pedro J Gonçalves, Rachel Rapp, Ismail Elezi, Stefanie Liebe, Jakob H Macke, Florian Mormann, Laura Leal-Taixé
Formal analysis: Franziska Gerken, Alana Darcher
Funding: Jakob H Macke, Florian Mormann, Laura Leal-Taixé
Software: Franziska Gerken, Alana Darcher, Johannes Niediek, Thomas P Reber
Project administration: Jakob H Macke, Florian Mormann, Laura Leal-Taixé
Supervision: Pedro J Gonçalves, Ismail Elezi, Stefanie Liebe, Jakob H Macke, Florian Mormann, Laura Leal-Taixé
Writing - original draft: Franziska Gerken, Alana Darcher
Writing - review editing: Pedro J Gonçalves, Rachel Rapp, Ismail Elezi, Stefanie Liebe, Jakob H Macke, Florian Mormann, Laura Leal-Taixé

References

[1] R. Q. Quiroga, L. Reddy, C. Koch, and I. Fried. Decoding visual inputs from multiple neurons in the human temporal lobe. 98 (4):1997–2007. ISSN 0022-3077. doi: 10.1152/jn.00125.2007. URL https://journals.physiology.org/doi/full/10.1152/jn.00125.2007. Publisher: American Physiological Society.

[2] R. Q. Quiroga, G. Kreiman, C. Koch, and I. Fried. Sparse but not 'grandmother-cell'coding in the medial temporal lobe. *Trends in cognitive sciences*, 12(3):87–91, 2008.

[3] H. G. Rey, M. J. Ison, C. Pedreira, A. Valentin, G. Alarcon, R. Selway, M. P. Richardson, and R. Quian Quiroga. Single-cell recordings in the human medial temporal lobe. *Journal of anatomy*, 227(4):394–408, 2015.

[4] G. Kreiman, C. Koch, and I. Fried. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature neuroscience*, 3(9):946–953, 2000.

[5] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

[6] F. Mormann, J. Niediek, O. Tudusciuc, C. M. Quesada, V. A. Coenen, C. E. Elger, and R. Adolphs. Neurons in the human amygdala encode face identity, but not gaze direction. *Nature neuroscience*, 18(11):1568–1570, 2015.

[7] F. Mormann, J. Dubois, S. Kornblith, M. Milosavljevic, M. Cerf, M. Ison, N. Tsuchiya, A. Kraskov, R. Q. Quiroga, R. Adolphs, et al. A category-specific response to animals in the right human amygdala. *Nature neuroscience*, 14(10):1247–1249, 2011.

[8] T. P. Reber, M. Bausch, S. Mackay, J. Boström, C. E. Elger, and F. Mormann. Representation of abstract semantic knowledge in populations of human single neurons in the medial temporal lobe. *PLOS Biology*, 17(6):1–17, 06 2019. doi: 10.1371/journal.pbio.3000290. URL https://doi.org/10.1371/journal.pbio.3000290.

[9] M. J. Ison, R. Quian Quiroga, and I. Fried. Rapid encoding of new memories by individual neurons in the human brain. 87(1):220–230. ISSN 0896-6273. doi: 10.1016/j.neuron.2015.06.016. URL https://www.sciencedirect.com/science/article/pii/S0896627315005590.

[10] J. Kamiński, S. Sullivan, J. M. Chung, I. B. Ross, A. N. Mamelak, and U. Rutishauser. Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nature neuroscience*, (4):590–601, 2017.

[11] N. A. Herweg, L. Kunz, D. Schonhaut, A. Brandt, P. A. Wanda, A. D. Sharan, M. R. Sperling, A. Schulze-Bonhage, and M. J. Kahana. A learned map for places and concepts in the human medial temporal lobe. *Journal of Neuroscience*, 43(19):3538–3547, 2023. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0181-22.2023. URL https://www.jneurosci.org/content/43/19/3538.

[12] J. Niediek, J. Boström, C. E. Elger, and F. Mormann. Reliable analysis of single-unit recordings from the human brain under noisy conditions: tracking neurons over hours. *PloS one*, 11(12):e0166598, 2016.

[13] J. M. Zacks, T. S. Braver, M. A. Sheridan, D. I. Donaldson, A. Z. Snyder, J. M. Ollinger, R. L. Buckner, and M. E. Raichle. Human brain activity time-locked to perceptual event boundaries. 4(6):651–655. ISSN 1546-1726. doi: 10.1038/88486. URL https://www.nature.com/articles/nn0601_651.

[14] C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, and K. A. Norman. Discovering event structure in continuous narrative perception and memory. 95(3):709–721.e5. ISSN 1097-4199. doi: 10.1016/j.neuron.2017.06.041.

[15] J. Zheng, A. G. Schjetnan, M. Yebra, B. A. Gomes, C. P. Mosher, S. K. Kalia, T. A. Valiante, A. N. Mamelak, G. Kreiman, and U. Rutishauser. Neurons detect cognitive boundaries to structure episodic memories in humans. *Nature neuroscience*, 25(3): 358–368, 2022.

[16] L. Isik, J. Singer, J. R. Madsen, N. Kanwisher, and G. Kreiman. What is changing when: Decoding visual information in movies from human intracranial recordings. *NeuroImage*, 180:147–159, 2018. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.08.027. URL https://www.sciencedirect.com/science/article/pii/S1053811917306742. New advances in encoding and decoding of brain signals.

[17] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach. Intersubject synchronization of cortical activity during natural vision. *science*, 303(5664):1634–1640, 2004.

[18] J. N. v. d. Meer, M. Breakspear, L. J. Chang, S. Sonkusare, and L. Cocchi. Movie viewing elicits rich and reliable brain state dynamics. *Nature communications*, 11(1):5004, 2020.

[19] Y. Zhang, Z. M. Aghajan, M. Ison, Q. Lu, H. Tang, G. Kalender, T. Monsoor, J. Zheng, G. Kreiman, V. Roychowdhury, et al. Decoding of human identity by computer vision and neuronal vision. *Scientific Reports*, 13(1):1–16, 2023.

[20] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

[21] Z. E. Markow, K. Tripathy, A. M. Svoboda, M. L. Schroeder, S. M. Rafferty, E. J. Richter, A. T. Eggebrecht, M. A. Anastasio, M. A. Chevillet, E. M. Mugler, S. N. Naufel, A. Yin, J. W. Trobaugh, and J. P. Culver. Identifying naturalistic movies from human brain activity with high-density diffuse optical tomography. *bioRxiv*, 2023. doi: 10.1101/2023.11.27.566650. URL https://www.biorxiv.org/content/early/2023/11/28/2023.11.27.566650.

[22] M. Khosla, G. H. Ngo, K. Jamison, A. Kuceyeski, and M. R. Sabuncu. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. 7(22):eabe7547. doi: 10.1126/sciadv.abe7547. URL https://www.science.org/doi/full/10.1126/sciadv.abe7547. Publisher: American Association for the Advancement of Science.

[23] L. Le, L. Ambrogioni, K. Seeliger, Y. Güçlütürk, M. v. Gerven, and U. Güçlü. Brain2pix: Fully convolutional naturalistic video reconstruction from brain activity. URL https://www.biorxiv.org/content/10.1101/2021.02.02.429430v2. Pages: 2021.02.02.429430 Section: New Results.

[24] StabilityAI. Stable diffusion. https://beta.dreamstudio.ai/generate, 2024.

[25] T. P. Reber, J. Faber, J. Niediek, J. Boström, C. E. Elger, and F. Mormann. Single-neuron correlates of conscious perception in the human medial temporal lobe. *Current Biology*, 27(19):2991–2998, 2017.

[26] S. Panzeri, J. H. Macke, J. Gross, and C. Kayser. Neural population coding: combining insights from microscopic and mass signals. *Trends in Cognitive Sciences*, 19(3):162–172, March 2015. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2015.01.002. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(15)00003-0. Publisher: Elsevier.

[27] E. Schneidman, J. L. Puchalla, R. Segev, R. A. Harris, W. Bialek, and M. J. Berry. Synergy from silence in a combinatorial neural code. *Journal of Neuroscience*, 31(44):15732–15741, 2011.

[28] C. J. Parde, Y. I. Colón, M. Q. Hill, C. D. Castillo, P. Dhar, and A. J. O'Toole. Closing the gap between single-unit and neural population codes: Insights from deep learning in face recognition. *Journal of Vision*, 21(8):15–15, 08 2021. ISSN 1534-7362. doi: 10.1167/jov.21.8.15. URL https://doi.org/10.1167/jov.21.8.15.

[29] M. Jamali, B. L. Grannan, E. Fedorenko, R. Saxe, R. Báez-Mendoza, and Z. M. Williams. Single-neuronal predictions of others' beliefs in humans. *Nature*, 591(7851):610–614, 2021.

[30] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

[31] I. Fried, K. A. MacDonald, and C. L. Wilson. Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. 18(5):753–765. ISSN 0896-6273. doi: 10.1016/S0896-6273(00)80315-3. URL https://www.sciencedirect.com/science/article/pii/S0896627300803153.

[32] R. Cao, J. Wang, C. Lin, E. De Falco, A. Peter, H. G. Rey, J. DiCarlo, A. Todorov, U. Rutishauser, X. Li, N. J. Brandmeir, and S. Wang. Feature-based encoding of face identity by single neurons in the human amygdala and hippocampus. URL http://biorxiv.org/lookup/doi/10.1101/2020.09.01.278283.

[33] M. Bausch, J. Niediek, T. P. Reber, S. Mackay, J. Boström, C. E. Elger, and F. Mormann. Concept neurons in the human medial temporal lobe flexibly represent abstract relations between concepts. *Nature communications*, 12(1):6164, 2021.

[34] J. A. Garcia-Lazaro, L. A. C. Belliveau, and N. A. Lesica. Independent Population Coding of Speech with Sub-Millisecond Precision. *Journal of Neuroscience*, 33(49):19362–19372, December 2013. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3711-13.2013. URL https://www.jneurosci.org/content/33/49/19362. Publisher: Society for Neuroscience Section: Articles.

[35] R. A. A. Ince, S. Panzeri, and C. Kayser. Neural Codes Formed by Small and Temporally Precise Populations in Auditory Cortex. *Journal of Neuroscience*, 33(46):18277–18287, November 2013. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.2631-13.2013. URL https://www.jneurosci.org/content/33/46/18277. Publisher: Society for Neuroscience Section: Articles.

[36] F. Mormann, S. Kornblith, M. Cerf, M. J. Ison, A. Kraskov, M. Tran, S. Knieling, R. Quian Quiroga, C. Koch, and I. Fried. Scene-selective coding by single neurons in the human parahippocampal cortex. *Proceedings of the National Academy of Sciences*, 114(5):1153–1158, 2017.

[37] G. Dehnen, M. S. Kehl, A. Darcher, T. T. Müller, J. H. Macke, V. Borger, R. Surges, and F. Mormann. Duplicate detection of spike events: A relevant problem in human single-unit recordings. *Brain Sciences*, 11(6):761, 2021.

[38] A. Darcher, R. Rapp, and T. Mueller. Epiphyte, May 2024. URL https://github.com/mackelab/epiphyte.

[39] FFmpeg, 2012. URL https://ffmpeg.org/.

[40] O. Aubert and Y. Prié. Advene: active reading through hypervideo. pages 235–244, 2005.

[41] T. Esler. Facenet pytorch. https://github.com/timesler/facenet-pytorch.git.

[42] B. Castellano. PySceneDetect. URL https://github.com/Breakthrough/PySceneDetect.

[43] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[44] R. J. SIMES. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 12 1986. ISSN 0006-3444. doi: 10.1093/biomet/73.3.751. URL https://doi.org/10.1093/biomet/73.3.751.

[45] E. Maris and R. Oostenveld. Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190, 2007.

[46] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. URL http://arxiv.org/abs/1412.6980. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[48] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[49] A. D. Ekstrom, M. J. Kahana, J. B. Caplan, T. A. Fields, E. A. Isham, E. L. Newman, and I. Fried. Cellular networks underlying human spatial navigation. *Nature*, 425(6954):184–188, 2003.

[50] J. F. Miller, I. Fried, N. Suthana, and J. Jacobs. Repeating spatial activations in human entorhinal cortex. *Current Biology*, 25(8): 1080–1085, 2015.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.