Two forms of uncertainty hierarchically shape lexical predictions during natural speech processing

Guilleminot P.[1*] and Morillon B.[1]

[1*]Aix Marseille Université, INSERM, INS, Institut de Neurosciences des Système, Marseille, 13005, France.

*Corresponding author(s). E-mail(s): pierre.guilleminot@univ-amu.fr;
Contributing authors: bnmorillon@gmail.com;

Abstract

Speech comprehension relies on contextual lexical predictions to optimize the processing of incoming words. As in other predictive paradigms, it is commonly assumed that prediction error, or surprisal, is weighted by a single summary statistic: uncertainty. Uncertainty reflects the overall unpredictability of the context and is typically quantified using Shannon entropy. However, for predictions involving many alternatives, as in speech, Shannon entropy is an ambiguous measure. We investigated alternative measures of uncertainty, namely the family of Rényi entropies, of which Shannon entropy is a special case. Using whole-brain intracranial EEG recordings in 33 epilepsy patients listening to natural, continuous speech, we show that, rather than Shannon entropy, contextual lexical uncertainty in speech is best captured by two extremes of the Rényi family. These reflect the probability of the most likely word and the spread across all plausible alternatives, which we term strength and dispersion, respectively. These representations are processed in distinct neural populations in a spatiotemporal hierarchy, with the encoding of dispersion preceding that of strength. Around each word onset, information flows between these clusters in a top-down and subsequently bottom-up sequence. Finally, they interact differentially with word surprisal during word processing. Our findings demonstrate that the brain encodes a multidimensional representation of uncertainty and reveal that multiple summary statistics may coexist in predictive processing. More broadly, this work introduces a flexible framework for characterizing the neural bases of uncertainty. It also highlights the potential mismatch between formal definitions of cognitive variables and their neural implementation.

Keywords: Speech, Human electrophysiology, Predictive Coding, Information Theory, Uncertainty

Introduction

To adapt to a dynamic environment, the brain relies on predictions generated by a continuously updating model of the world. This framework, referred to as predictive coding, offers a probabilistic account of how the brain constantly minimizes the discrepancy between expected and actual input via the computation of prediction errors (Rao and Ballard, 1999; Friston, 2005; Heilbron and Chait, 2018). During speech perception, these predictive mechanisms are crucial to anticipate upcoming words throughout an ever changing context (Kalikow et al., 1977; Lau et al., 2008, 2009). Recent advances in large language models (LLMs) offer a promising avenue for studying this process. Trained on vast corpora, LLMs generate context-sensitive probability distributions over upcoming words, which are aligned with neural signals (Schrimpf et al., 2021; Caucheteux and King, 2022; Caucheteux et al., 2023). As such, LLMs provide computational observer models for speech, as commonly used in other statistical sequence learning tasks (Meyniel, 2020; Badre et al., 2012; Stern et al., 2010).

Two key computational quantities derived from these probability distributions are uncertainty, the amount of unknown information about the next input (Shannon, 1951), and surprisal, the amount of information actually conveyed by that input (Shannon, 1948). Both are represented in the human brain (Frank et al., 2015; Meyniel et al., 2015; Wilmes et al., 2023). The cortex appears to track uncertainty continuously, updating it with each incoming unit of surprisal, which in turn modulates the integration of future inputs (Stern et al., 2010; Donhauser and Baillet, 2020; Meyniel, 2020; Heilbron et al., 2022). Traditionally, uncertainty over discrete probability distributions, such as lexicon-bounded predictions, is quantified using Shannon entropy, a foundational measure of unpredictability grounded in information theory (Shannon, 1948) and widely adopted in neuroscience (Donhauser and Baillet, 2020; Friston, 2010; Bach and Dolan, 2012; Garner, 1962; Heilbron et al., 2022; Hasson, 2017). However, despite its utility, Shannon entropy compresses a distribution into a single statistic, conflating qualitatively different types of uncertainty. For instance, whether uncertainty arises from the number of possible alternatives (0.12, 0.12, 0.76) or from the absence of a clear most likely event (.5, .5, 0) (Churchland et al., 2008; Yeon and Rahnev, 2020) is not captured by Shannon's entropy (H = 1.0) (Fig6A,B,D,E). Whether the brain distinguishes between these distinct forms of uncertainty remains unknown.

To address this, we use a generalization of Shannon entropy: the Rényi entropy family (Rényi, 1961). The latter introduces a parameter, the Renyi order α, that modulates the weight given to unlikely events, thereby offering several measures of uncertainty. Low α values emphasize the number of plausible events i.e. events with a non-null probability regardless of its actual value, a form of uncertainty we term *dispersion*. On the contrary, high α values prioritize the probability of the most likely event, which we term *strength*. While these aspects are indistinguishable in binary outcomes, they become dissociable in distributions containing more possible events

(N≥3). Although Rényi entropies have been applied in fields such as computer science and EEG signal analysis (Baraniuk et al., 2002; Sahoo and Arora, 2004; Liang et al., 2015; Müller-Lennert et al., 2013), they have yet to be used to characterize the neural representation of uncertainty during naturalistic cognition.

Here, we capitalized on information-theoretic tools and human intracranial recordings to examine whether the brain represents multiple forms of uncertainty during naturalistic speech processing, and how these would interact with incoming words. Our results reveal the coexistence of two separate neural populations encoding different forms of contextual uncertainty: one reflecting dispersion across plausible alternatives, and another capturing the strength of dominant prediction. This finding shows that the brain does not rely on Shannon entropy, nor on the large set of Rényi entropy family, but instead maintains two complementary representations of uncertainty. Critically, these representations intervene at different stages of the speech processing hierarchy: Early sensory processing relies on pre-stimulus dispersion in primary auditory areas, possibly to optimally process expected words, as expected by a Bayesian framework. On the contrary, higher semantic processing possibly relies on post-stimuli strength which is encoded in a distributed network to increase the weight of unexpected inputs, as expected by an entropy reduction framework. Tracking the flow of uncertainty-related information transfer between these two stages of the cortical hierarchy further revealed that pre-stimulus, top-down anticipatory signals allowed to prepare for incoming stimuli, while post-stimulus sensory signals were propagated bottom-up. Speech processing thus relies on the communication between different stages of the cortical hierarchy, each with their own forms of uncertainty that serve complementary functions. The coexistence of dispersion and strength therefore provides a general solution for summarizing probability distributions in complex tasks without tracking the probability of every possible alternative.

Results

The brain does not represent the Shannon entropy

To characterize whether the human brain encodes different measures of lexical uncertainty during speech comprehension, we recorded intracranial EEG from 33 epileptic patients as they listened to 10 mn of an audiobook (Fig.1A). Data were re-referenced using a bipolar montage (Mercier et al., 2022) and epoched to word onsets. Using the transcript, we generated word-by-word probability distributions with CamemBERT (Martin et al., 2019), a large language model fine-tuned for French. From these, we computed 20 Rényi entropies for each word (Rényi, 1961), spanning α values from 0.01 to 10 (Fig.1A). Because neural coding of probability is often non-linear (Foucault et al., 2024), we used Gaussian Copula Mutual Information (GCMI) to relate neural data to entropy values (Fig.1A), a rank-based non-parametric estimator suitable to estimate non-linear, monotonic relationships in low-data regimes (Ince et al., 2017). Around 9 percent of channels (456/5126) showed statistically significant mutual information with at least one type of Rényi entropy (1000 permutations, $p = 0.05$, cluster-corrected over time).

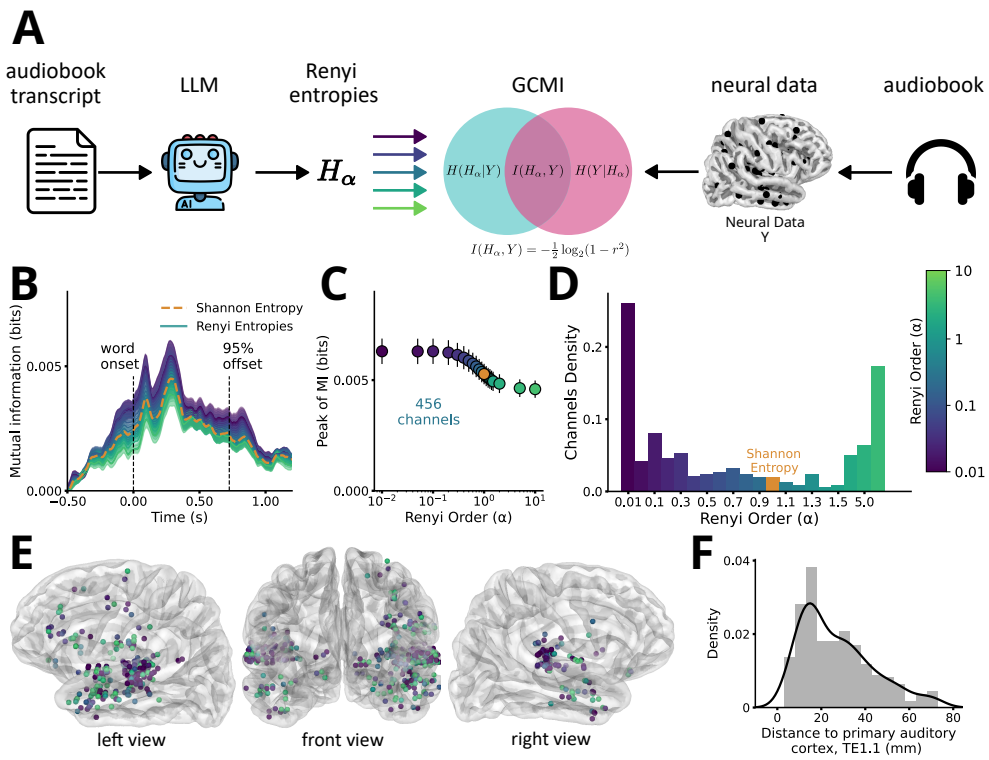


Fig. 1 Neural representation of lexical uncertainty (A) Analysis pipeline. Left: Transcript of an audiobook processed through an LLM to extract word probability distributions at each word onset. Renyi entropies (H_α) are computed for each distribution. Right: Intracranial EEG recorded during audiobook listening, epoched to word onset. Middle: MI between Renyi entropies and epoched neural data, computed across words using Gaussian copula (GC) at each timestep between −200 ms and 1200 ms. **(B)** Time course of Mutual information (MI) between LLM-based Renyi entropies and neural responses to speech, averaged across 456 significant channels for different Renyi orders (α, color gradient). **(C)** MI peak as a function of Renyi order, averaged across channels. **(D)** Distribution across channels of the Renyi order yielding maximum MI with neural data. **(E)** Cortical distribution of channels as a function of the Renyi order yielding maximal MI. **(F)** Histogram of distances from each channel to primary auditory cortex (posteromedial Heschl's gyrus, TE1.1).

To assess the temporal dynamics of uncertainty encoding, we computed mutual information (MI) timecourses grand-averaged across participants and channels. This analysis revealed 3 prominent MI peaks, at respectively −50 ms pre-onset, 100 ms post onset and 280 ms post onset, with sustained activity extending up to 1 s. Both peaks and sustained activity were modulated by the Rényi order α (Fig.1B). To quantify the dependence on α, we extracted the most prominent peak MI amplitude for each α value, showing that Shannon entropy ($\alpha = 1$) is not the most strongly represented measure, contrary to conventional assumptions (Fig.1C). Instead, on average across channels, entropies with lower α values, emphasizing distributional dispersion, consistently yielded higher MI with neural data. The relationship followed a sigmoidal profile

(inflection point at $\alpha = 0.78$, plateau below $\alpha = 0.1$). This pattern suggests that overall neural activity preferentially tracks the dispersion of possible outcomes over the strength of the most likely prediction. To evaluate the generalizability of this result, we repeated the analysis at the individual level (Fig.9B), finding that 31 of 33 participants exhibited a similar profile, with only 2 displaying a general preference for higher-α measures. Across subjects, inflection points were located at $\alpha = 0.84 \pm 0.39$ (Fig.9A-C).

We next examined whether this preference for low-α entropies was consistent at the level of individual recording sites. For each channel, we identified the α value yielding the highest MI peak, which revealed a bimodal distribution rather than a single preferred uncertainty measure (Fig.1D). Specifically, most significant channels encoded information about either the strength (high α) or the dispersion (low α) of the distribution, but not Shannon entropy ($\alpha = 1$). Because strength reflects the informational content of the most likely event, we added to our analysis an additional measure: the informational content of the second most likely event. This feature displayed no statistically significant mutual information with the neural data, thus hinting that the brain does not track the probability of multiple events individually, but rather computes the most likely probability as a summary statistics (Fig.8.A). Using a searchlight procedure, we next showed that the information relative to uncertainty is encoded across distributed cortical regions, with the different types of entropy being broadly represented, thus hinting at a spatial code (Fig.1F-G)

Together, these results point towards the multidimensional representation of lexical uncertainty, indicating that Shannon entropy can be an ill-representation of uncertainty in contexts with multiple possible outcomes. The neural representation of uncertainty during naturalistic speech processing is instead built upon the two extremes of the Renyi family: the uncertainty relative to the most likely word (strength) and the uncertainty relative to the number of likely (non-null) candidates (dispersion).

Different neural populations encodes different uncertainties

To characterize the heterogeneity in preferred entropy measures across channels, we applied non-negative matrix factorization (NMF) as a soft clustering method (Yu et al., 2024). Clustering was performed on the time-averaged MI values from word onset to the 95% word offset, across α for all statistically significant channels (N=456; Fig.8B). This analysis revealed a simple two-cluster structure, each corresponding to a distinct uncertainty representation. One cluster (248 channels present in all 33 participants) exhibited maximal sensitivity to low-α Rényi values, thus to dispersion (Fig.2A-C). A second cluster (164 channels present in all 33 participants) exhibited maximal sensitivity to high α values, thus to prediction strength (Fig.2D–F). For clarity, these are hereafter referred to as the dispersion cluster and the strength cluster, respectively.

The temporal dynamics of both clusters differs. The dispersion cluster displays all 3 peaks of the overall MI timecourse (Fig1B), at -50 ms, 100 ms and 280 ms relative to word onset together with sustained activity extending up to 1 s (Fig.2A). In contrast, the strength cluster displays a single peak at 280 ms, that remains elevated and decays gradually up to 1 s (Fig.2D). To quantify the α-dependence within each cluster, we extracted the peak MI amplitudes as a function of α. The dispersion cluster exhibits

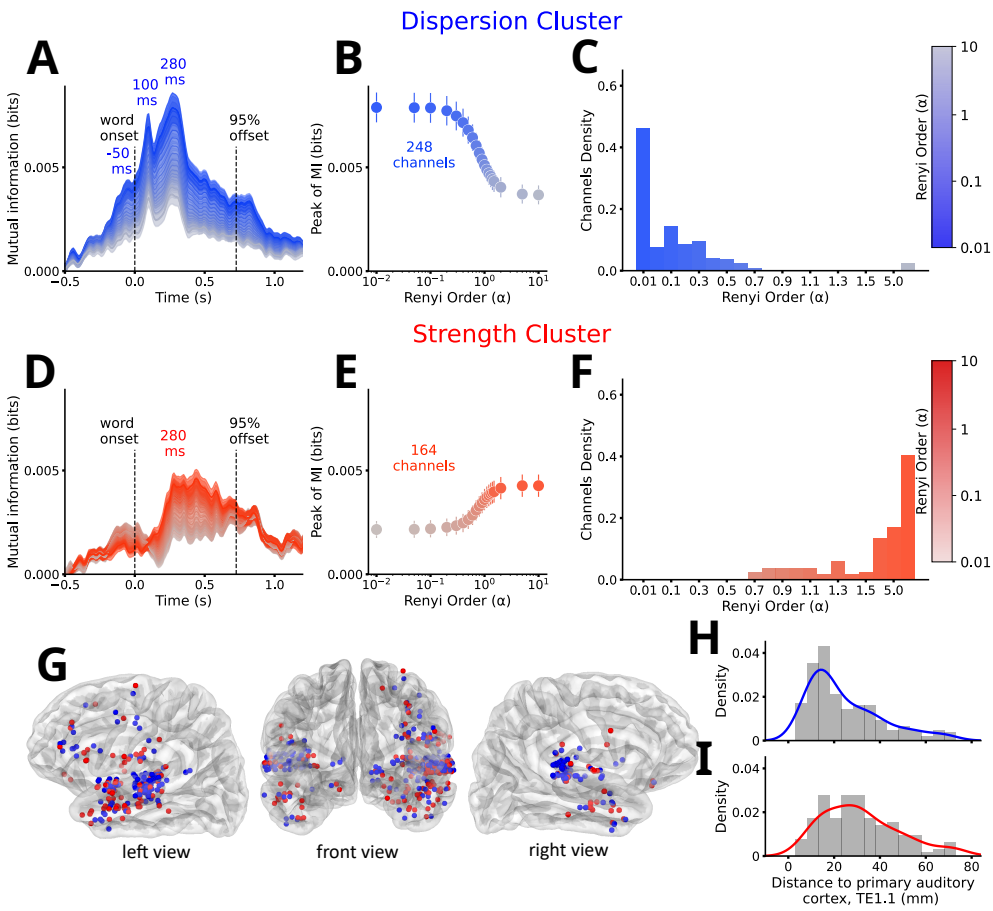


Fig. 2 Dispersion and Strength clusters of channels. (A-C) Dispersion cluster ($n = 248$ channels). **(A)** Grand average of MI over time across channels for different Renyi orders (α, color gradient). **(B)** MI peak as a function of Renyi order, averaged across channels. **(C)** Histogram of Renyi order yielding maximum MI with neural data, across channels. **(D-F)** Strength cluster ($n = 164$ channels). Same conventions as A-C. **(G)** Cortical distribution of the Dispersion (blue) and Strength (Red) clusters. **(H,I)** Histogram of distance to primary auditory cortex (posteromedial Heschl's gyrus, TE1.1) across channels in the (**H**) Dispersion and (**I**) Strength clusters.

a sigmoidal profile with an inflection point at $\alpha = 0.70$, plateauing near $\alpha = 0.1$, whereas the strength cluster shows a sigmoidal dependence with an inflection point at $\alpha = 0.82$, plateauing near $\alpha = 5$ (Fig.2B,E).

Examining the distribution of preferred α values across channels (Fig.2C,F) confirmed that the clustering efficiently captures the bimodal split of MI peaks (Fig.1D). Moreover, the two clusters display minimal overlap across α values, indicating that they encode qualitatively different measures of uncertainty rather than forming a

continuous gradient. Individual-level analyses showed that all participants have comparable profiles, with subject-wise inflection points at $\alpha = 0.68 \pm 0.14$ for the dispersion cluster and $\alpha = 0.84 \pm 0.29$ for the strength cluster (Fig.9D-I).

Finally, spatial representation and euclidean distance from primary auditory cortex (Norman-Haignere et al., 2024) showed a partially dissociated anatomical distribution (two-tailed t-test, t = -4.01, $df = 340$, $p < 10^{-4}$) (Fig.2G-I). The dispersion cluster is predominantly localized in peri-auditory lower-level sensory areas, and the strength cluster more in associative cortical regions. This spatial dissociation suggests that different representations of uncertainty are relevant at different stages of the speech processing cortical hierarchy.

Dispersion and strength present distinct temporal dynamics within their respective cluster

To understand the roles of dispersion and strength in shaping neural responses to word processing, it is necessary to disentangle their respective contribution to the neural signal. As these features display high mutual information between themselves (Fig.7A-C), we capitalized on Partial Information Decomposition (PID) (Williams and Beer, 2010). This analytic tool allows to decompose the joint MI between two features and a third variable (neural activity) into unique, redundant and synergistic components (Daube et al., 2019) (Fig.3A). PID revealed that the dispersion cluster exclusively encodes unique information relative to its preferred uncertainty measure (dispersion), presenting sustained pre-onset activity, followed by a first early peak at 100 ms post-onset and a second peak at 300 ms post-onset (ttests: $p < 0.05$ FDR-corrected for multiple comparisons). On the contrary, the strength cluster shows a small unique contribution of dispersion at the word onset, and a large unique contribution of strength between 100 ms and 900 ms post word onset (ttests: $p < 0.05$, FDR-corrected for multiple comparisons) (Fig.3C,D). Both clusters present high redundant information, in line with the high mutual information shared between strength and dispersion (Fig.7A-C), with timecourses that closely resemble those of the unique information in each cluster (Fig.3E). Finally, none of the two clusters exhibit synergistic interaction between strength and dispersion, hinting at a parallel processing of the two summary statistics in distinct channels (Fig.3F). Overall, these results reveal a well-separated encoding of strength and dispersion within their respective clusters.

Given that uncertainty and surprisal share mutual information (Fig7A-C) and are thought to interact in response to incoming stimuli, we next investigated their respective contributions to neural activity. We thus applied the PID between uncertainty (dispersion or strength, in their respective clusters), surprisal and neural activity (Fig.3B). This analysis first confirmed that dispersion uniquely contributes to neural activity pre-word onset, together with the early post-onset activity at around 100 ms, while strength is only encoded later, from 200 ms post-onset onward (Fig.3G). In contrast, activity peaking around 280 ms is in fact uniquely contributed by surprisal in both clusters (Fig.3H). The unique surprisal at word onset on the strength cluster can be attributed to the shared information between dispersion and surprisal and thus reflect the unique dispersion activity found on the strength cluster (Fig.3C).

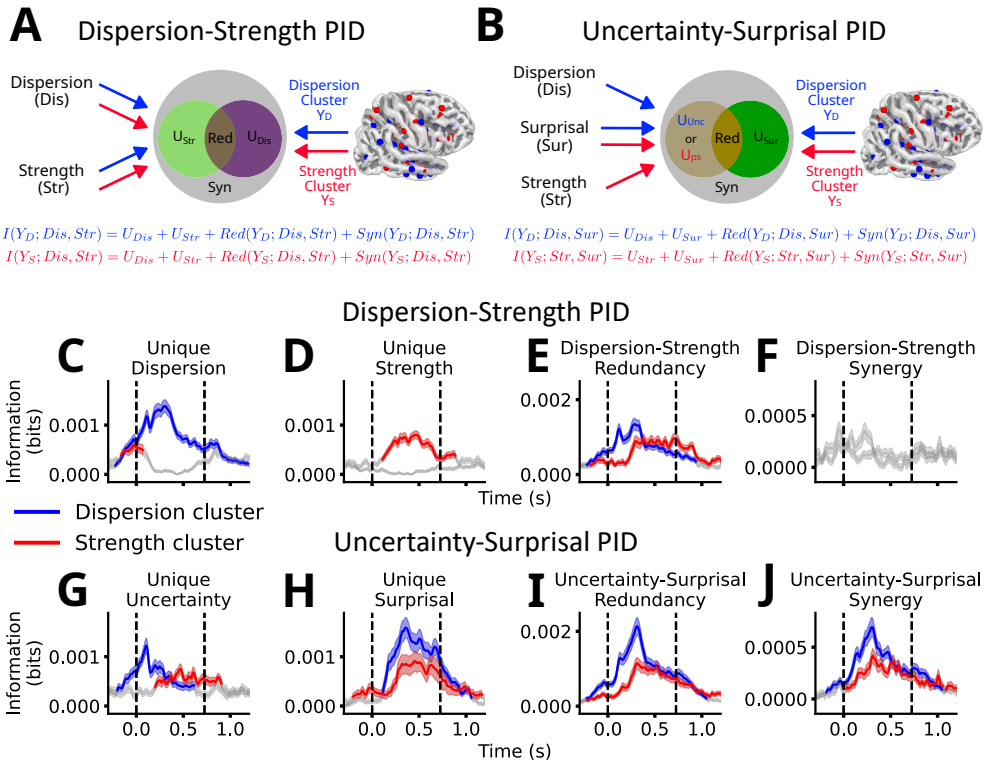


Fig. 3 Partial Information Decomposition (PID) of the joint mutual information between neural data and LLM features. (A) Dispersion-Strength PID analysis framework: For each channel cluster (dispersion, blue; strength, red), the joint mutual information between neural data, dispersion (Renyi entropy with $\alpha = 0.01$) and strength (Renyi entropy with $\alpha = 10$) is computed. This is decomposed into the information about the neural data carried uniquely by dispersion, uniquely by strength, redundantly by both, and synergistically when combined. **(B)** Uncertainty-Surprisal PID analysis framework: For each channel cluster, the joint mutual information between neural data, surprisal (the unexpectedness of the observed word given the context) and the preferred uncertainty ($\alpha = 0.01$ or $\alpha = 10$) for that cluster is computed. **(C,D,E,F)** Timecourses of the Dispersion-Strength PID components. **(G,H,I,J)** Timecourses of the Uncertainty-Surprisal PID components. Bold lines represent statistical significance compared to baseline (ttests: $p < 0.05$, FDR-corrected for multiple comparisons).

In addition to validating the previous analyses regarding uncertainty processing (Fig.3G), as well as showing post-onset sustained response to surprisal in both clusters (Fig.3H), we observed both redundant and synergistic information between uncertainty and surprisal (Fig.3I,J). In both cases, activity peaks earlier in the dispersion cluster (at 100 ms and 280 ms post-onset) compared to the strength cluster (single 280 ms peak). Overall, these results highlight the different timecourses of the encoding of dispersion (early) and strength (late) on their respective clusters and the similar interaction between uncertainty and surprisal in both clusters.

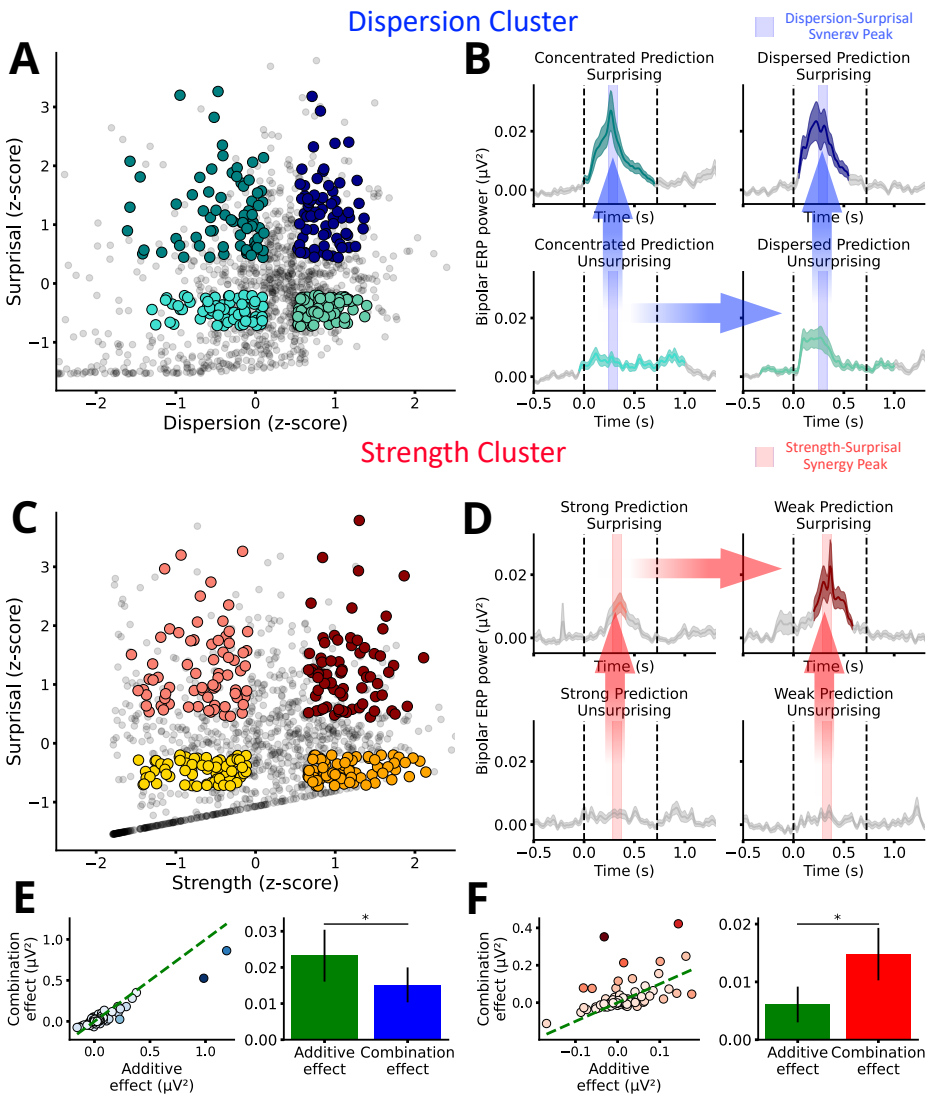


Fig. 4 Evoked responses in the the Dispersion and Strength channel clusters as a function of surprisal and uncertainty. (A-B) Dispersion cluster. **(A)** Division of word epochs as a function of surprisal and dispersion. **(B)** Power ERPs, zscored relative to baseline, of the different quadrants in the Dispersion cluster. **(C-D)** Strength cluster. **(C)** Division of word epochs as a function of surprisal and strength. **(D)** Power ERPs, zscored relative to baseline, of the different quadrants in the Strength cluster. Colored lines correspond to statistically relevant activity compared to a −1000 ms to −500 ms baseline. Transparent blue/red vertical areas correspond to the time window of maximal uncertainty-surprisal synergy, as defined by the PID analysis (Fig.3J). Blue/red arrows represent significant increases in power within the synergy window between two adjacent quadrants. **(E,F)** Effect of the combination of uncertainty and surprisal, compared to an additive model of these features (green), for the dispersion ((**E**)) and strength ((**F**)) clusters.

Dispersion and strength interact respectively sub-additively and supra-additively with surprisal

To further investigate the synergistic interactions between uncertainty and surprisal (highlighted by the PID analyses), we performed ERP-based analyses. Across words, we orthogonalized the contributions of uncertainty and surprisal, to assess their individual and joint effects on neural activity, in both clusters (Fig.4A,C). This resulted in 4 orthogonal categories (quadrants) of words: low-uncertainty + low-surprisal (bottom left), low-uncertainty + high surprisal (top left), high-uncertainty + low-surprisal (bottom right), and high-uncertainty + high surprisal (top right; all $bf < 0.17$, Fig.11).

In the dispersion cluster, significant ERPs power is observed for all 4 quadrants ($p < 0.01$, FDR corrected). Both dispersion and surprisal individually contribute to the power of the response, in addition to a significant effect when combined (Fig.4B, colored arrows). Specifically, dispersion only yielded an effect when the stimulus was not surprising ($p < 0.005$, FDR corrected). In contrast surprisal yielded an effect whether the stimuli predictions were dispersed or not ($p < 0.001$, FDR corrected). To further characterize the nature of the synergy between dispersion and surprisal, we computed for each channel an additive model as the sum of the effect of surprisal in the absence of dispersion (from lower-left to upper-left), and of dispersion in the absence of surprisal (from lower-left to lower-right). This additive model was then compared to the combined effect of surprisal and dispersion (from lower-left to upper-right) (Fig.4E). This analysis revealed that the combined effect of surprisal and dispersion was sub-additive ($p = 0.005$), indicating that within early sensory areas, uncertainty dampens surprising inputs, accordingly to a Bayesian sensory framework.

In sharp contrast, the strength cluster significantly responded only for highly-surprising words ($p < 0.01$, FDR corrected), indicating that it does not respond to all sensory inputs, but exclusively to prediction errors. Surprisal yielded an effect whether the stimulus predictions were strong or not ($p < 0.005$, FDR corrected). However, contrary to the dispersion cluster, strength only yielded an effect when the stimuli were surprising ($p < 0.005$, FDR corrected) and the combined effect of surprisal and strength was supra-additive ($p < 0.05$; Fig.4F). This shows that within these higher-level associative areas, uncertainty amplifies the representation of surprisal, accordingly to a predictive coding framework.

Together, these results show that dispersion and strength are represented in neural clusters of activity that hold different functions within the speech processing hierarchy. Specifically, dispersion is computed in early sensory areas and acts as a prior to optimally integrate incoming inputs, while strength operates in higher-level areas where it positively modulates surprisal to increase information gain.

Uncertainty information is transmitted top-down and then bottom-up

To characterize the transfer of information between the two clusters, we finally combined an information theory framework with a connectivity approach. The feature-specific information transfer (FIT) provides such a metric as it allows to identify the transfer of information between two neural channels that is specific to a feature of

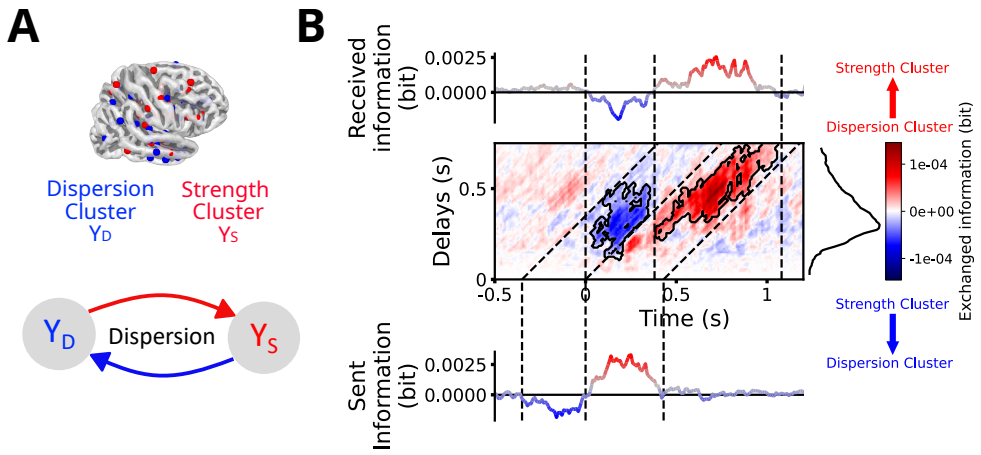


Fig. 5 Feature-specific information transfer (FIT) relative to dispersion between the dispersion and strength clusters. (A) Dispersion FIT framework: for each pair of channels across clusters within the same subject, the net information transfer of dispersion is computed. **(B)** Time-delay FIT map. The transferred information was summed across delays to compute the received information (top), and summed across diagonals to compute the sent information (bottom). The absolute value of information was averaged across times to compute the average delay between sender and receiver (right). Statistical significance in the FIT map is displayed using contours. Negative values (blue) correspond to a transfer from the strength cluster to the dispersion one, while positive values (red) correspond to a transfer from the dispersion cluster to the strength one.

interest. The application of the FIT results in a time-delay matrix, containing the amount of information being transmitted between the two clusters at a certain time and from a certain delay in the past, with negative and positive values representing the direction of the information flow. This metric was computed over all possible pairs of channels from the two clusters within each subject separately, resulting in 956 pairs.

Computing this metric over strength, dispersion and surprisal features revealed similar patterns (Fig.9), due to the shared information between them (Fig.10B), with the clearest effect for dispersion (Fig.5A). Statistical analysis across all pairs of channels on the time-delay matrix revealed two significant clusters of information transfer (permutation cluster approach, $p < 1^{-5}$,Fig.5B). The information was first sent from the strength cluster around -150 ms pre-onset to the dispersion cluster around 100 ms post-onset. It was then sent back from the dispersion cluster around 200 ms post-word onset to the strength cluster around 600 ms post-word onset. On average, the delay of information transfer between the two clusters was 260 ms (Fig.5B, right insert).

These results demonstrate a clear bidirectional transfer of information between two stages of the speech processing hierarchy. First, higher-level areas, which encode the prediction strength of the most likely upcoming word, send pre-onset anticipatory information to sensory areas (top-down information flow) as to prepare for incoming stimulus. Then, sensory areas, which encode the dispersion across all plausible words,

send back information related to stimuli processing to higher level areas (bottom-up information flow).

Discussion

While Shannon entropy is the most widely used metric for quantifying uncertainty in both computational and cognitive neurosciences, our findings show that the brain does not represent lexical uncertainty as such. Rather, it relies on two complementary measures of uncertainty, reflecting 1- the probability of the most likely event and 2- the number of plausible events, which we termed strength and dispersion. These two forms of uncertainty have specific functions and are represented at distinct stages of the speech processing hierarchy. Early sensory processing in the auditory cortex relies predominantly on pre-stimulus dispersion, an anticipatory signal that facilitates the processing of expected stimuli, with uncertainty dampening surprising inputs, according to a Bayesian framework. On the contrary, later lexical processing in distributed downstream networks relies on prediction strength to amplify unexpected inputs, thus improving the gain of surprisal information according to a predictive coding framework. Moreover, the cortical clusters associated to the encoding of dispersion and strength communicate during speech processing: higher-level areas send anticipatory information to lower-level areas (top-down) to prepare for incoming signals, which in turns send sensory information to higher-level areas for processing (bottom-up). Together, these results point towards a two-stages hierarchical organization of speech processing in which two complementary forms of lexical uncertainty are represented to provide a general solution for summarizing probability distributions without tracking the probability of every possible word..

When multiple alternatives are possible, the brain relies on of a summary statistics rather than on the whole probability distribution, thus leading to suboptimalities (Churchland et al., 2008; Yeon and Rahnev, 2020; Ma and Jazayeri, 2014). The nature of the summary statistics can be ambiguous for complex distributions and tasks (Daikoku, 2018), such as language processing. Here we addressed this problematic by parameterizing uncertainty using the Renyi entropy family, whose members discount low-probability events to varying degrees, represented by a single parameter α ranging from 0 to $+\infty$ (Crupi et al., 2018). Our results demonstrate that during speech comprehension, lexical uncertainty is represented as the two extremes of this family: the uncertainty relative to the number of all possible words ($\alpha \to 0$) and the uncertainty of the most probable word ($\alpha \to \infty$). The former, which we termed dispersion, converges to the Hartley entropy $H_0(X)$ (Wehrl, 1990), though in practice we computed it as $H_{0.01}(X)$ to accommodate for the fact that the probability output of large language models is never null. Dispersion represents the number of plausible next words for a given context and is therefore analogous to neighborhood size and lexical cohort density De Cara and Goswami (2003); Kowialiewski and Majerus (2019). In contrast, when $\alpha \to \infty$, the entropy only reflects the probability of the most likely word. Here, this form of uncertainty was termed strength, as it is defined relative to the the strength of the prediction, but is also found in the literature as the Bayes error, error entropy or min-entropy (Fano and Hawkins, 1961). Contrary to

dispersion, strength relies fully on the notion of probability and focuses on a single event, and is therefore particularly meaningful for prediction tasks Yeon and Rahnev (2020). Critically, we show that the neural representation of prediction strength is limited to the most likely upcoming word, as shown by the absence of significant encoding of the probability of the second most likely word (Fig.9A). Thus, we propose that in complex tasks relying on highly dimensional probability distribution, the neural implementation of predictive coding relies on a combination of two forms of uncertainty: dispersion and strength. What is considered as highly dimensional for human brains remains to be addressed in future work.

We show that the two forms of uncertainty are processed by distinct neural populations characterized by different spatio-temporal dynamics. Dispersion-sensitive channels were predominantly localized in early cortical auditory areas and exhibited pre-word activity, likely reflecting the representation of prior auditory templates (Kok et al., 2017), followed by an early post-word onset activity, analogous to the N100 and thus related to sensory processing Schafer and Marcus (1973); Näätänen and Picton (1987). In contrast, strength-sensitive channels were more distributed, predominantly in associative higher-order regions. They only encoded prediction strength after 200 ms post-stimulus onset when unexpected words were perceived, which can be related to the N400 and higher-order lexical processing (Tyler et al., 2005; Fedorenko et al., 2012; Frank et al., 2015; Yang et al., 2024). Importantly, both populations displayed strong synergistic interaction between uncertainty and surprisal at a latency peaking around 280 ms, linking them to surprisal processing and context updating (Donchin and Coles, 1988). We further show that in addition to holding different functions within the cortical hierarchy, these cortical populations interact with each other, in a dynamic that reflects their coordinated role in speech processing. Strength-sensitive channels, associated to higher level processes, inform dispersion-sensitive populations in early sensory regions pre-stimuli. The latter then send bottom-up information, possibly about the actual input, in the opposite direction (Heilbron et al., 2022). This alternative top-down and bottom-up information flow allows the brain to adapt to a continuous influx of words. The spatial, temporal and inter-clusters dynamics thus all point towards dispersion and strength being associated to different, dynamically interacting stages of the speech processing hierarchy.

Our work ties into a growing body of work distinguishing uncertainty and surprisal in sensory neuroscience (Kok et al., 2012; De Lange et al., 2018; Teufel et al., 2018), and especially during language comprehension (Frank et al., 2015; Willems et al., 2016; Donhauser and Baillet, 2020; Song et al., 2024). Crucially, both forms of uncertainty synergistically interact with surprisal within their respective stage of the speech processing hierarchy. In dispersion-sensitive channels, associated to early sensory processing, the combination of surprisal and dispersion was sub-additive. This suggests that uncertainty dampens the response to surprising words, consistent with predictive coding theory, where precision positively modulates prediction errors in sensory systems Friston (2010); Meyniel (2020). On the contrary, in strength-sensitive channels, associated to higher-level processes, the combination of surprisal and strength was supra-additive. This suggests uncertainty amplified the response to surprising words,

consistent with mechanisms aiming to reduce contextual uncertainty, i.e. increase informational gain in predictive systems (Roesch et al., 2012; Richter et al., 2018; Song et al., 2024). Both forms of uncertainty thus modulate the response to surprising inputs, but are associated to different neural processes within their respective hierarchical stage, accordingly to recent views (Press et al., 2020) and hypothesized neural mechanisms (Angela and Dayan, 2005).

Although our study focused on lexical predictions, language processing operates across multiple representational levels, and it is likely that different linguistic units (phrases, syllables, phonemes) are associated with distinct entropy profiles (Donhauser and Baillet, 2020; Gibson et al., 2019). Furthermore, given their functional dissociation, strength and dispersion, as well as the information they exchange, may be associated to different frequency-specific dynamics (Bastos et al., 2015; Fontolan et al., 2014). While our findings are specific to speech processing, the proposed framework of parameterizing uncertainty into distinct measures may generalize to other complex prediction tasks. We thus advocate for a broader use of generalized entropy families (Crupi et al., 2018; Sharma and Mittal, 1975) to accurately evaluate the representation of uncertainty in neural data.

Materials and methods

Data and code availability

The raw data investigated in the current manuscript is privileged patient data. Therefore the conditions of our ethics approval do not permit public archiving of anonymised data. Readers seeking access to the data should contact Dr. Daniele Schön (daniele.schon@univ-amu.fr). Access will be granted to named individuals in accordance with ethical procedures governing the reuse of clinical data, including completion of a formal data sharing agreement. Data analyses were performed using custom scripts in Python, which are available on GitHub : https://github.com/phg17/witching-star.

Participants

sEEG recordings were obtained from 33 patients with refractory epilepsy using intracerebral electrodes implanted as part of the standard presurgical evaluation process . Electrode placements were based solely on clinical requirements. All patients were French native speakers. Neuropsychological assessments carried out before sEEG recordings indicated that they all had intact language functions and met the criteria for normal hearing. None of them had their epileptogenic zone including the auditory areas as identified by experienced epileptologists. Recordings took place at the Hôpital de La Timone (Marseille, France). All patients gave informed consent, and the experiment reported here was approved by the Institutional Review board of the French Institute of Health .

sEEG recordings

The sEEG signal was recorded using depth electrodes shafts of 0.8-mm diameter containing 10 to 15 electrode contacts (Dixi Medical or Alcis, Besançon, France). The contacts were 2 mm long and were spaced from each other by 1.5 mm. Data were recorded using a 256 -channels amplifier (Brain Products), sampled at 1000 Hz and high-pass filtered at 0.016 Hz inside a soundproof insulated Faraday cage. The recording reference and ground were chosen by the clinical staff as two consecutive sEEG contacts on the same shaft both located in the white matter and/or at distance from any epileptic activity. Neural recordings were performed between 3 and 10 days after the implantation procedure. No sedation or analgesics drugs were used, and antiepileptic drugs were partially or completely withdrawn. Recordings were always acquired more than 4 hours after the last seizure. Patients were included in the study if their implantation map covered at least partially the Heschl's gyrus (left or right).

Preprocessing

To increase spatial sensitivity and reduce passive volume conduction from neighboring regions (Mercier et al. (2017, 2022)), the signal was offline re-referenced using bipolar montage. That is, for a pair of adjacent electrode contacts, the referencing led to a virtual channel located at the midpoint locations of the original contacts. To precisely localize the channels, a procedure similar to the one used in the iELVis toolbox and in the FieldTrip toolbox was applied (Groppe et al. (2017); Stolk et al. (2018)). First, we manually identified the location of each channel centroid on the post-implant CT scan using the Gardel software (Villalon et al. (2018)). Second, we performed volumetric segmentation and cortical reconstruction on the pre-implant MRI with the Freesurfer image analysis suite (documented and freely available for download online at http://surfer.nmr.mgh.harvard.edu/). This segmentation of the pre-implant MRI with SPM12 provides us with both the tissue probability maps (i.e. gray, white, and cerebrospinal fluid [CSF] probabilities) and the indexed-binary representations (i.e. either gray, white, CSF, bone, or soft tissues). This information allowed us to reject electrodes not located in the brain. Third, the post-implant CT scan was coregistered to the pre-implant MRI via a rigid affine transformation and the pre-implant MRI was registered to MNI152 space, via a linear and a non-linear transformation from SPM12 methods (Penny et al. (2011)), through the FieldTrip toolbox (Oostenveld et al. (2011)). Fourth, applying the corresponding transformations, we mapped channel locations to the pre-implant MRI brain that was labeled using the volume-based Human Brainnetome Atlas (Fan et al. (2016)).

Based on the brain segmentation performed using SPM12 methods through the FieldTrip toolbox, bipolar channels located outside of the brain were removed from the data (3%). The remaining data was then bandpass filtered between 0.1 and 250 Hz, and, following a visual inspection of the power spectral density profile of the data, when necessary, we additionally applied a notch filter at 50 Hz and harmonics up to 200 Hz to remove power line artifacts . Finally, the data were downsampled to 100 Hz.

Experimental Design

The experiment consisted in a single session in which patients passively listened to 10 min of storytelling 577 s, La sorcière de la rue Mouffetard (Gripari and Dahl (1994)).

In the Faraday cage, a sound Blaster X-Fi Xtreme Audio, an amplifier Yamaha P2040 and Yamaha loudspeakers (NS 10M) were used for sound presentation. In the bedroom, stimuli were presented using a Sennheiser HD 25 headphone set. Sound stimuli were presented at 44.1 kHz sample rate and 16 bits resolution. Speech and music excerpts were presented at 75 dBA.

Computations of predictions via a Large Language Model

The neural recordings were compared with the predictions of a pretrained large language model (LLM) (Radford et al. (2019)) in response to the same story as presented to subjects. Because the stories were in French, we used the Camembert model (Martin et al. (2019)), a performant language model based on RoBERTa (Liu et al. (2019)) and fined-tuned for French. We used the pretrained models from Huggingface. For each word of the story, the model used all previous words as context to make predictions about the next one. These predictions take the form of a probability distribution over the 32k subword tokens.

Computation of surprisal and uncertainty

Given the distribution of possible tokens, we computed the surprisal of the first subtoken of each observed word as:

$$I(X) = -\log(p_x) \tag{1}$$

$$\tag{2}$$

with p_x the probability of the observed token.

Traditionally, uncertainty is computed as the Shannon entropy over the distribution of possible events (tokens) i.e. the expected value of surprisal:

$$H_{Shannon}(X) = -\sum_{x \in \mathcal{X}} p_x log(p_x) \tag{3}$$

with $\mathcal{X}$ the distribution of tokens. While this formulation captures the variety of distribution shapes for two events (Fig.**??**B), it becomes ambiguous for more events (Fig.**??**E). Here, we sought to refine the formulation of uncertainty, using a generalization of the Shannon entropy: the Renyi entropy (Rényi (1961)). The latter relies on a weaker postulates about the additive property of entropy. This allows to extend the Shannon entropy to a family of quantities, characterized by their Renyi order α, as:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log\left(\sum_{x \in \mathcal{X}} p_x^\alpha\right) \tag{4}$$

This newly defined quantity allows to give different weights to unlikely events (Fig.**??**I), as they lose importance with higher α. As such, when $\alpha \to 0$, all plausible events (events with non-null probability) are considered equally, this case is referred to as the Hartley entropy or dispersion. On the contrary, when $\alpha \to +\infty$, only the probability of the most likely event is considered. Hence, every intermediary value of α represents different trade-offs between dispersion (the number of plausible events) and strength (the probability of the most likely event) (Fig.**??**K). This quantity further covers the classical Shannon entropy as $H_{\alpha \to 1} = H_{Shannon}$. Uncertainty is thus not defined by a single quantity, but rather by an entropy profile, which allows to distinguish between different distribution shapes (Fig.**??**C,F,J). Hence, given the probability distribution of all possible tokens at every word onset, the uncertainty was computed as the Renyi entropy for 20 values of α: 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 5 and 10. We have approximated Dispersion by $\alpha = 0.01$ since the output of the language never outputs exactly null probabilities and the Hartley entropy is thus be constant across the dataset (Klir, 2006).

Gaussian Copula Mutual Information

To quantify the statistical dependency between neural data and each of our 20 measures of uncertainty, we used an information-based measure: the mutual information (MI), defined as:

$$I(X;Y) = H(X) - H(X|Y) \tag{5}$$

where X and Y represent the broadband neural data epoched over words and the uncertainties at each of these, with $H(X)$ and $H(Y)$ their respective entropy. $H(X|Y)$ is the conditional entropy of X given Y. More specifically, we adopted the Gaussian Copula Mutual Information (GCMI) (Ince et al. (2017)), which does not require binning to compute MI. It is a robust rank-based approach that allows to detect non-linear monotonic relation between variables and is a lower-bound approximate estimator of MI. Before using this framework, we first transformed all variables (features and neural data) into Gaussian variables using the inverse normal transformation. The MI was then computed as:

$$MI(X;Y) = -\frac{1}{2ln(2)} \frac{ln(|\Sigma_X||\Sigma_Y|)}{|\Sigma_{XY}|} \tag{6}$$

with Σ_X, Σ_Y and Σ_{XY} respectively being the covariance matrices of X, Y and of the joint variable (X, Y). The GCMI was computed at each time-point of the epochs, for every channel. We thus obtained a time-series for each channel and for each uncertainty. All computations were performed using the frites toolbox (Combrisson et al. (2022))

We thus obtained timecourses of mutual information for each channel and each value of α. From these, we extracted the value at the peak of the averaged timecourse

for each channel and value of alpha, resulting in a single value per channel and per Renyi order. This allowed us to plot the entropy profile (Fig1C, Fig2B,E).

Non-negative Matrix Factorisation

To characterize the heterogeneity in preferred entropy measures across channels, we applied non-negative matrix factorization (NMF) as a soft clustering method, appropriate given the strictly positive nature of mutual information (MI) (Yu et al. (2024); Lee and Seung (1999)). The NMF iteratively approximates a nonnegative data matrix $\mathbf{X} \in \mathcal{R}^{N,p}$ by:

$$\mathbf{X} \approx \mathbf{WH} \tag{7}$$

where $\mathbf{W} \in \mathcal{R}^{N,r}$ and $\mathbf{H} \in \mathcal{R}^{r,p}$ are nonnegative matrices respectively referred to as the basis matrix and the coefficient matrix, with $r \leq max(N, p)$ the number of components. The coefficient matrix $\mathbf{H}$ represents how strongly a data point belongs to each of the component and can therefore be interpreted as a membership matrix (Li and Ding (2018)). Using a threshold on the latter, we can obtain a soft clustering of the data. For each time series previously obtained, we computed the peak of activity, thus obtaining 20 strictly positive data points per channel, on which the NMF was applied. Hyperparameters (number of clusters, regularization parameters) were optimized using the silhouette scoreShahapure and Nicholas (2020). Channels with a soft membership below 0.65 were excluded to retain only those with strong cluster membership.

Partial Information Decomposition (PID)

The goal of studying different definition of entropies is to be able to identify the one that best explain the neural data. However, this task is non-trivial as different entropies are highly correlated. Therefore, it is necessary to be able to dissociate the unique and redundant information they contain about the neural data. Moreover, one essential aspect of predictive coding is the interaction between prediction(uncertainty) and the observed signal (the surprisal). To address these two questions, we employed the Partial Information Decomposition (PID) (Williams and Beer (2010); Wibral et al. (2015)) framework. The latter decompose the joint mutual information that two variables X_1 and X_2 hold about a third variable Y, into 4 information atoms: the atoms uniquely attributed to respectively X_1 and X_2, the atom redundantly attributed to both and the atom synergistically created when observed together:

$$I(X_1, X_2; Y) = U(X_1; Y) + U(X_2; Y) + R(X_1, X_2; Y) + S(X_1, X_2; Y) \tag{8}$$

To calculate this decomposition, we first computed the redundancy measure as the minimum of the information provided by each individual feature about the target (Barrett (2015)):

$$R(X_1, X_2; Y) = min\left[I(X_1, Y), I(X_2, Y)\right] \tag{9}$$

The unique information terms can then be computed algebraically as:

$$U(X_1;Y) = MI(X_1;Y) - R(X_1, X_2;Y) \tag{10}$$
$$U(X_2;Y) = MI(X_2;Y) - R(X_1, X_2;Y) \tag{11}$$

And finally synergy is given by:

$$S(X_1, X_2;Y) = I(X_1, X_2;Y) - MI(X_1;Y) - MI(X_2;Y) + R(X_1, X_2;Y) \tag{12}$$

This approach is most commonly used to study how the combination of different neural signal informs about a single feature(Combrisson et al. (2024); Greco et al. (2024)). In our case, it is more fit to quantify the contribution of different features on a single channel (Daube et al. (2019)). Hence, the unique information allows to determine the specificity to which a channel respond to one feature over another, while redundancy and synergy informs us about the interactions between them Ince et al. (2017); Kay and Ince (2018). All computations were performed using the frites toolbox (Combrisson et al. (2022)).

For each component of the PID and each cluster, we used the information from -500 ms to -250 ms pre-onset as a baseline to establish statistical relevancy of the information between -250 ms and 1250 ms relative to word onset.

Feature-specific Information Transfer (FIT)

To characterize the transfer of information between two channels, Y_1 and Y_2, about a feature, X, it is necessary to combine an information theory approach with Granger causality principle. To achieve that, we used the feature-specific information transfer (FIT). This method identifies the information about a feature, X a channel, $Y_{2,pres}$ has in common with the past of another channel, $Y_{1,past}$, while controlling for its own past, $Y_{2,past}$, Celotto et al. (2023); Lemke et al. (2024). It relies on the PID introduced earlier and can be formulated as follow:

$$FIT = min[SUI(X : Y_{1,past}, Y_{2,pres} \backslash Y_{2,past}), SUI(Y_{2,pres} : Y_{1,past}, X \backslash Y_{2,past})] \tag{13}$$

where $SUI(X : Y_{1,past}, Y_{2,pres} \backslash Y_{2,past})$ is the information shared by $Y_{1,past}$ and $Y_{2,present}$ about X that is not contained in $Y_{2,past}$. $SUI(Y_{2,pres} : Y_{1,past}, X \backslash Y_{2,past})$ is the information shared by $Y_{2,pres}$ and X about $Y_{1,past}$ that is not contained in $Y_{1,past}$. As such, $FIT(Y_1 \rightarrow Y_2; X)$ quantifies the amount of information about the feature, X, transferred from Y_1 to Y_2.

Here the FIT was used to assess the directionality of the information transfer between our two clusters, relative to our different information features. To do so, we computed both $FIT(Y_1 \rightarrow Y_2; X)$ and $FIT(Y_2 \rightarrow Y_1; X)$ and subtracted one from the other for each lag and channel. This resulted in a time-delay matrix, containing the difference of information being transmitted between channels at a certain time and from a certain delay in the past. Negative and positive values thus signifies opposite direction for the information transfer. This metric was computed over all possible pairs of channels from both clusters within each subject separately, resulting in 956 pairs.

The FIT was computed for delays going up to 1 second. Statistical significance of time-delay regions was assessed using permutation clusters with 1000 permutations accross pairs of channels. By averaging the matrix across delays, we computed the timecourse of average received information. Then, by averaging data along time-delay diagonal, we obtained the timecourse of average sent information. Finally, by averaging along timepoints, we could obtain the average delay of communication between clusters.

Orthogonalisation of Surprisal and Uncertainty

In order to characterize the interaction between uncertainty and surprisal within each cluster, we separated the words in our corpus according to their surprisal and the uncertainty of their associated context. This procedure was done separately for each form of uncertainty revealed by previous analysis: the Renyi entropies for $\alpha = 0.01$ and $\alpha = 10$. First we represented each word within the corpus as a point in a 2D space, with the x-axis representing uncertainty and the y-axis surprisal. Then, within that space, we selected the points that fell within the first and last 30% for the surprisal and within the within the first and last 45% for the uncertainty. These values were chosen as to keep at least 10% of the total number of words within each of the quarters we thus defined (Fig.4A,C). Then using a procedural methods, we extracted from each square, 5% of the total number of words as to minimize the Bayes factor of the distribution of the data projected on each axis, for their respective neighboring quarter Kass and Raftery (1995). This resulted in 4 orthogonal quarters of words: low-uncertainty + low-surprisal (bottom left), low-uncertain + high surprisal (top left), high-uncertainty + low-surprisal (bottom right), high-uncertain + high surprisal, for both dispersion and strength (all $bf < 0.17$, Fig.12).

ERP power analysis

We computed the ERP power for each of the 4 categories, relative to the cluster and uncertainty of interest, using the -1000 ms to -500 ms activity pre-word onset as a baseline to assess statisical relevance. The ERP power was computed as we are using a bipolar montage and therefore, polarity cannot be interpreted. Comparison between quadrants was done based on a 100 ms window around the peaks of synergy between uncertainty and surprisal highlighted by the PID: 300 ms post-onset for the dispersion cluster and 320 ms post-onset for strength cluster. Effects of the different features were computed as the difference between the synergy region in a quadrant compared to an orthogonal quadrant (Fig.4B,D, colored arrows). Non-linearity was assessed as the difference between the sum of individual effects of surprisal (from bottom-left to top-left) and uncertainty (from bottom-left to bottom-right), and the combination of them (from bottom-left to top-right). The statistical significance of this difference was then assessed using a ttest.

Spatial representation

Because of the sparsity of intracranial implantation, spatial representations do not generalize well. Here we ran a procedure, akin to a searchlight procedure, in order to delete spatially isolated channels, or groups of channels driven by a single subject.

When spatially representing our data(Fig1E, Fig2G), we first considered the neighborhood of each point as a 10 mm sphere around them, and only plotted channels that contained a channel from a different subject within their neighborhood.

Moreover, in order to compare the spatial representation of each cluster of activity, we reduced their position within the 3D space as their distance to the primary auditory cortex, as te1.1 Kiviniemi et al. (2009); Anderson et al. (2011); Norman-Haignere et al. (2024). Their distances could then be compared using a ttest.

References

Rao, R. P.; Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* **1999**, *2*, 79–87.

Friston, K. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences* **2005**, *360*, 815–836.

Heilbron, M.; Chait, M. Great expectations: is there evidence for predictive coding in auditory cortex? *Neuroscience* **2018**, *389*, 54–73.

Kalikow, D. N.; Stevens, K. N.; Elliott, L. L. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the acoustical society of America* **1977**, *61*, 1337–1351.

Lau, E. F.; Phillips, C.; Poeppel, D. A cortical network for semantics:(de) constructing the N400. *Nature reviews neuroscience* **2008**, *9*, 920–933.

Lau, E.; Almeida, D.; Hines, P. C.; Poeppel, D. A lexical basis for N400 context effects: Evidence from MEG. *Brain and language* **2009**, *111*, 161–172.

Schrimpf, M.; Blank, I. A.; Tuckute, G.; Kauf, C.; Hosseini, E. A.; Kanwisher, N.; Tenenbaum, J. B.; Fedorenko, E. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2105646118.

Caucheteux, C.; King, J.-R. Brains and algorithms partially converge in natural language processing. *Communications biology* **2022**, *5*, 134.

Caucheteux, C.; Gramfort, A.; King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour* **2023**, *7*, 430–441.

Meyniel, F. Brain dynamics for confidence-weighted learning. *PLOS Computational Biology* **2020**, *16*, e1007935.

Badre, D.; Doll, B. B.; Long, N. M.; Frank, M. J. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron* **2012**, *73*, 595–607.

Stern, E. R.; Gonzalez, R.; Welsh, R. C.; Taylor, S. F. Updating beliefs for a decision: neural correlates of uncertainty and underconfidence. *Journal of Neuroscience* **2010**, *30*, 8032–8041.

Shannon, C. E. Prediction and entropy of printed English. *Bell system technical journal* **1951**, *30*, 50–64.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal* **1948**, *27*, 379–423.

Frank, S. L.; Otten, L. J.; Galli, G.; Vigliocco, G. The ERP response to the amount of information conveyed by words in sentences. *Brain and language* **2015**, *140*, 1–11.

Meyniel, F.; Sigman, M.; Mainen, Z. F. Confidence as Bayesian probability: From neural origins to behavior. *Neuron* **2015**, *88*, 78–92.

Wilmes, K. A.; Petrovici, M. A.; Sachidhanandam, S.; Senn, W. Uncertainty-modulated prediction errors in cortical microcircuits. *bioRxiv* **2023**, 2023–05.

Donhauser, P. W.; Baillet, S. Two distinct neural timescales for predictive speech processing. *Neuron* **2020**, *105*, 385–393.

Heilbron, M.; Armeni, K.; Schoffelen, J.-M.; Hagoort, P.; De Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2201968119.

Friston, K. The free-energy principle: a unified brain theory? *Nature reviews neuroscience* **2010**, *11*, 127–138.

Bach, D. R.; Dolan, R. J. Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature reviews neuroscience* **2012**, *13*, 572–586.

Garner, W. R. Uncertainty and structure as psychological concepts. **1962**,

Hasson, U. The neurobiology of uncertainty: implications for statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2017**, *372*, 20160048.

Churchland, A. K.; Kiani, R.; Shadlen, M. N. Decision-making with multiple alternatives. *Nature neuroscience* **2008**, *11*, 693–702.

Yeon, J.; Rahnev, D. The suboptimality of perceptual decision making with multiple alternatives. *Nature communications* **2020**, *11*, 3857.

Rényi, A. On measures of entropy and information. Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics. 1961; pp 547–562.

Baraniuk, R. G.; Flandrin, P.; Janssen, A. J.; Michel, O. J. Measuring time-frequency information content using the Rényi entropies. *IEEE Transactions on Information theory* **2002**, *47*, 1391–1409.

Sahoo, P. K.; Arora, G. A thresholding method based on two-dimensional Renyi's entropy. *Pattern Recognition* **2004**, *37*, 1149–1161.

Liang, Z.; Wang, Y.; Sun, X.; Li, D.; Voss, L. J.; Sleigh, J. W.; Hagihira, S.; Li, X. EEG entropy measures in anesthesia. *Frontiers in computational neuroscience* **2015**, *9*, 16.

Müller-Lennert, M.; Dupuis, F.; Szehr, O.; Fehr, S.; Tomamichel, M. On quantum Rényi entropies: A new generalization and some properties. *Journal of Mathematical Physics* **2013**, *54*.

Mercier, M. R.; Dubarry, A.-S.; Tadel, F.; Avanzini, P.; Axmacher, N.; Cellier, D.; Del Vecchio, M.; Hamilton, L. S.; Hermes, D.; Kahana, M. J.; others Advances in human intracranial electroencephalography research, guidelines and good practices. *Neuroimage* **2022**, *260*, 119438.

Martin, L.; Muller, B.; Suárez, P. J. O.; Dupont, Y.; Romary, L.; de La Clergerie, É. V.; Seddah, D.; Sagot, B. CamemBERT: a tasty French language model. *arXiv preprint arXiv:1911.03894* **2019**,

Foucault, C.; Bounmy, T.; Demortain, S.; Thiririon, B.; Eger, E.; Meyniel, F. A nonlinear code for event probability in the human brain. *BioRxiv* **2024**, 2024–02.

Ince, R. A.; Giordano, B. L.; Kayser, C.; Rousselet, G. A.; Gross, J.; Schyns, P. G. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human brain mapping* **2017**, *38*, 1541–1573.

Yu, L.; Dugan, P.; Doyle, W.; Devinsky, O.; Friedman, D.; Flinker, A. A left-lateralized dorsolateral prefrontal network for naming. *bioRxiv* **2024**,

Norman-Haignere, S. V.; Keshishian, M. K.; Devinsky, O.; Doyle, W.; McKhann, G. M.; Schevon, C. A.; Flinker, A.; Mesgarani, N. Temporal integration in human auditory cortex is predominantly yoked to absolute time, not structure duration. *bioRxiv* **2024**,

Williams, P. L.; Beer, R. D. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515* **2010**,

Daube, C.; Ince, R. A.; Gross, J. Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Current Biology* **2019**, *29*, 1924–1937.

Ma, W. J.; Jazayeri, M. Neural coding of uncertainty and probability. *Annual review of neuroscience* **2014**, *37*, 205–220.

Daikoku, T. Neurophysiological markers of statistical learning in music and language: Hierarchy, entropy and uncertainty. *Brain sciences* **2018**, *8*, 114.

Crupi, V.; Nelson, J. D.; Meder, B.; Cevolani, G.; Tentori, K. Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search. *Cognitive science* **2018**, *42*, 1410–1456.

Wehrl, A. *The many facets of entropy*; 1990.

De Cara, B.; Goswami, U. Phonological neighbourhood density: Effects in a rhyme awareness task in five-year-old children. *Journal of child language* **2003**, *30*, 695–710.

Kowialiewski, B.; Majerus, S. Verbal working memory and linguistic long-term memory: Exploring the lexical cohort effect. *Memory & Cognition* **2019**, *47*, 997–1011.

Fano, R. M.; Hawkins, D. Transmission of information: A statistical theory of communications. *American Journal of Physics* **1961**, *29*, 793–794.

Kok, P.; Mostert, P.; De Lange, F. P. Prior expectations induce prestimulus sensory templates. *Proceedings of the National Academy of Sciences* **2017**, *114*, 10473–10478.

Schafer, E. W.; Marcus, M. M. Self-stimulation alters human sensory brain responses. *Science* **1973**, *181*, 175–177.

Näätänen, R.; Picton, T. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* **1987**, *24*, 375–425.

Tyler, L. K.; Marslen-Wilson, W. D.; Stamatakis, E. A. Differentiating lexical form, meaning, and structure in the neural language system. *Proceedings of the National Academy of Sciences* **2005**, *102*, 8375–8380.

Fedorenko, E.; Nieto-Castanon, A.; Kanwisher, N. Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia* **2012**, *50*, 499–513.

Yang, Y.; Li, L.; de Deyne, S.; Li, B.; Wang, J.; Cai, Q. Unraveling lexical semantics in the brain: Comparing internal, external, and hybrid language models. *Human*

Brain Mapping **2024**, *45*, e26546.
Donchin, E.; Coles, M. G. Is the P300 component a manifestation of context updating? *Behavioral and brain sciences* **1988**, *11*, 357–374.
Kok, P.; Jehee, J. F.; De Lange, F. P. Less is more: expectation sharpens representations in the primary visual cortex. *Neuron* **2012**, *75*, 265–270.
De Lange, F. P.; Heilbron, M.; Kok, P. How do expectations shape perception? *Trends in cognitive sciences* **2018**, *22*, 764–779.
Teufel, C.; Dakin, S. C.; Fletcher, P. C. Prior object-knowledge sharpens properties of early visual feature-detectors. *Scientific reports* **2018**, *8*, 10853.
Willems, R. M.; Frank, S. L.; Nijhof, A. D.; Hagoort, P.; Van den Bosch, A. Prediction during natural language comprehension. *Cerebral cortex* **2016**, *26*, 2506–2516.
Song, M.; Wang, J.; Cai, Q. The unique contribution of uncertainty reduction during naturalistic language comprehension. *Cortex* **2024**, *181*, 12–25.
Roesch, M. R.; Esber, G. R.; Li, J.; Daw, N. D.; Schoenbaum, G. Surprise! Neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain. *European Journal of Neuroscience* **2012**, *35*, 1190–1200.
Richter, D.; Ekman, M.; de Lange, F. P. Suppressed sensory response to predictable object stimuli throughout the ventral visual stream. *Journal of Neuroscience* **2018**, *38*, 7452–7461.
Press, C.; Kok, P.; Yon, D. The perceptual prediction paradox. *Trends in cognitive sciences* **2020**, *24*, 13–24.
Angela, J. Y.; Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **2005**, *46*, 681–692.
Gibson, E.; Futrell, R.; Piantadosi, S. P.; Dautriche, I.; Mahowald, K.; Bergen, L.; Levy, R. How efficiency shapes human language. *Trends in cognitive sciences* **2019**, *23*, 389–407.
Bastos, A. M.; Vezoli, J.; Bosman, C. A.; Schoffelen, J.-M.; Oostenveld, R.; Dowdall, J. R.; De Weerd, P.; Kennedy, H.; Fries, P. Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron* **2015**, *85*, 390–401.
Fontolan, L.; Morillon, B.; Liegeois-Chauvel, C.; Giraud, A.-L. The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nature communications* **2014**, *5*, 4694.
Sharma, B. D.; Mittal, D. P. New non-additive measures of entropy for discrete probability distributions. *J. Math. Sci* **1975**, *10*, 28–40.
Mercier, M. R.; Bickel, S.; Megevand, P.; Groppe, D. M.; Schroeder, C. E.; Mehta, A. D.; Lado, F. A. Evaluation of cortical local field potential diffusion in stereotactic electro-encephalography recordings: a glimpse on white matter signal. *Neuroimage* **2017**, *147*, 219–232.
Groppe, D. M.; Bickel, S.; Dykstra, A. R.; Wang, X.; Mégevand, P.; Mercier, M. R.; Lado, F. A.; Mehta, A. D.; Honey, C. J. iELVis: An open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. *Journal of neuroscience methods* **2017**, *281*, 40–48.
Stolk, A.; Griffin, S.; Van Der Meij, R.; Dewar, C.; Saez, I.; Lin, J. J.; Piantoni, G.; Schoffelen, J.-M.; Knight, R. T.; Oostenveld, R. Integrated analysis of anatomical and electrophysiological human intracranial data. *Nature protocols* **2018**, *13*, 1699–1723.

Villalon, S. M.; Paz, R.; Roehri, N.; Lagarde, S.; Pizzo, F.; Colombet, B.; Bartolomei, F.; Carron, R.; Bénar, C.-G. EpiTools, A software suite for presurgical brain mapping in epilepsy: Intracerebral EEG. *Journal of neuroscience methods* **2018**, *303*, 7–15.

Penny, W. D.; Friston, K. J.; Ashburner, J. T.; Kiebel, S. J.; Nichols, T. E. *Statistical parametric mapping: the analysis of functional brain images*; Elsevier, 2011.

Oostenveld, R.; Fries, P.; Maris, E.; Schoffelen, J.-M. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience* **2011**, *2011*, 156869.

Fan, L.; Li, H.; Zhuo, J.; Zhang, Y.; Wang, J.; Chen, L.; Yang, Z.; Chu, C.; Xie, S.; Laird, A. R.; others The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral cortex* **2016**, *26*, 3508–3526.

Gripari, P.; Dahl, V. B. *La sorcière de la rue Mouffetard*; Systime, 1994.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; others Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**,

Klir, G. J. Uncertainty and information: foundations of generalized information theory. *Kybernetes* **2006**, *35*, 1297–1299.

Combrisson, E.; Basanisi, R.; Cordeiro, V. L.; Ince, R. A.; Brovelli, A. Frites: A Python package for functional connectivity analysis and group-level statistics of neurophysiological data. *Journal of open source software* **2022**, *7*, 3842.

Lee, D. D.; Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *nature* **1999**, *401*, 788–791.

Li, T.; Ding, C.-c. Nonnegative matrix factorizations for clustering: A survey. *Data Clustering* **2018**, 149–176.

Shahapure, K. R.; Nicholas, C. Cluster quality analysis using silhouette score. 2020 IEEE 7th international conference on data science and advanced analytics (DSAA). 2020; pp 747–748.

Wibral, M.; Lizier, J. T.; Priesemann, V. Bits from brains for biologically inspired computing. *Frontiers in Robotics and AI* **2015**, *2*, 5.

Barrett, A. B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Physical Review E* **2015**, *91*, 052802.

Combrisson, E.; Basanisi, R.; Neri, M.; Auzias, G.; Petri, G.; Marinazzo, D.; Panzeri, S.; Brovelli, A. Higher-order and distributed synergistic functional interactions encode information gain in goal-directed learning. *bioRxiv* **2024**, 2024–09.

Greco, A.; Moser, J.; Preissl, H.; Siegel, M. Predictive learning shapes the representational geometry of the human brain. *Nature Communications* **2024**, *15*, 9670.

Kay, J. W.; Ince, R. A. Exact partial information decompositions for Gaussian systems based on dependency constraints. *Entropy* **2018**, *20*, 240.

Celotto, M.; Bím, J.; Tlaie, A.; De Feo, V.; Toso, A.; Lemke, S.; Chicharro, D.; Nili, H.; Bieler, M.; Hanganu-Opatz, I.; others An information-theoretic quantification of the content of communication between brain regions. *Advances in Neural Information*

Processing Systems **2023**, *36*, 64213–64265.

Lemke, S. M.; Celotto, M.; Maffulli, R.; Ganguly, K.; Panzeri, S. Information flow between motor cortex and striatum reverses during skill learning. *Current Biology* **2024**, *34*, 1831–1843.

Kass, R. E.; Raftery, A. E. Bayes factors. *Journal of the american statistical association* **1995**, *90*, 773–795.

Kiviniemi, V.; Starck, T.; Remes, J.; Long, X.; Nikkinen, J.; Haapea, M.; Veijola, J.; Moilanen, I.; Isohanni, M.; Zang, Y.-F.; others Functional segmentation of the brain cortex using high model order group PICA. *Human brain mapping* **2009**, *30*, 3865–3886.

Anderson, J. S.; Ferguson, M. A.; Lopez-Larson, M.; Yurgelun-Todd, D. Reproducibility of single-subject functional connectivity measurements. *American journal of neuroradiology* **2011**, *32*, 548–555.

Supplementary information.

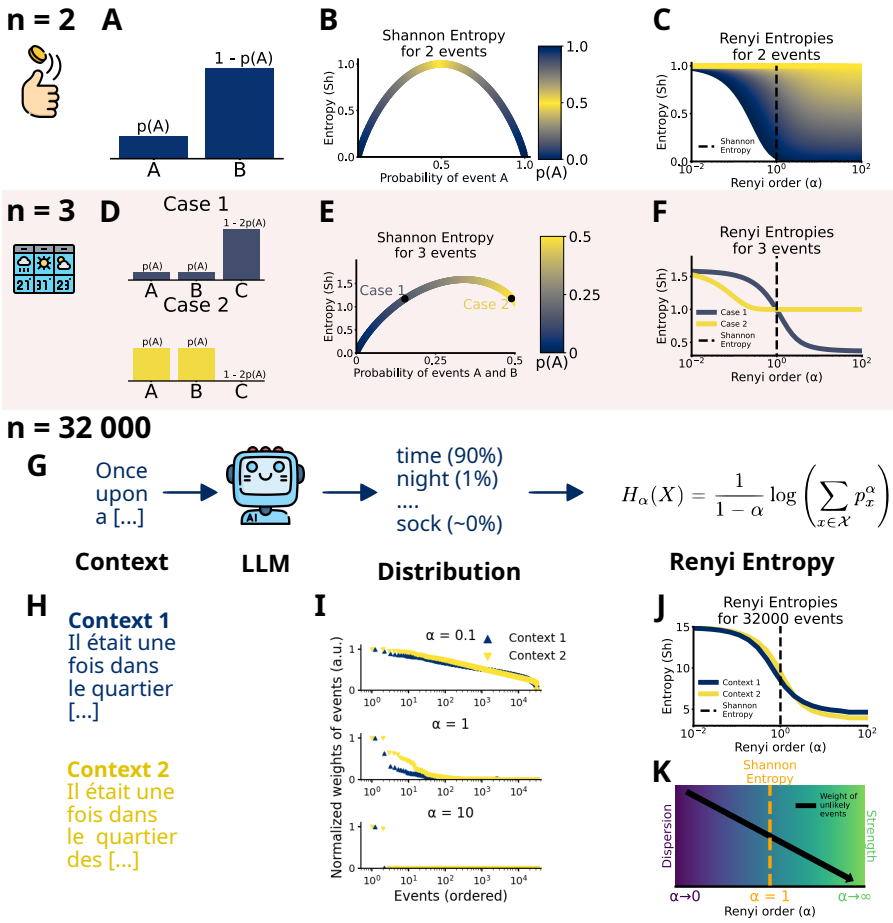


Fig. 6 Shannon and Renyi entropies for different distribution sizes. (A) A 2-event distribution, fully described by p(A). **(B)** Shannon entropy for $n = 2$ as a function of p(A). **(C)** Renyi entropy as a function of the Renyi order α, for different p(A) (color gradient). $\alpha = 1$ corresponds to Shannon entropy (dotted line). **(D)** Two cases of 3-event distributions with two equiprobable events, fully described by p(A). **(E)** Shannon entropy for $n = 3$ as a function of p(A) (= p(B), color gradient). **(F)** Renyi entropy as a function of α, for cases 1 and 2, which have the same Shannon entropy (dotted line). **(G)** In Large Language Models (LLMs), contextual probabilities are assigned to all possible words ($n = 32000$ events). **(H)** Two contexts example. **(I)** Normalized weights of events for different α. **(J)** Renyi entropy as a function of α, for contexts 1 and 2. **(K)** The entropies describing the most likely event ($\alpha \to \infty$, green) and all non-null events ($\alpha \to 0$, night blue) are respectively referred to as the Prediction Strength (Str) and the Distribution Dispersion (Dis).

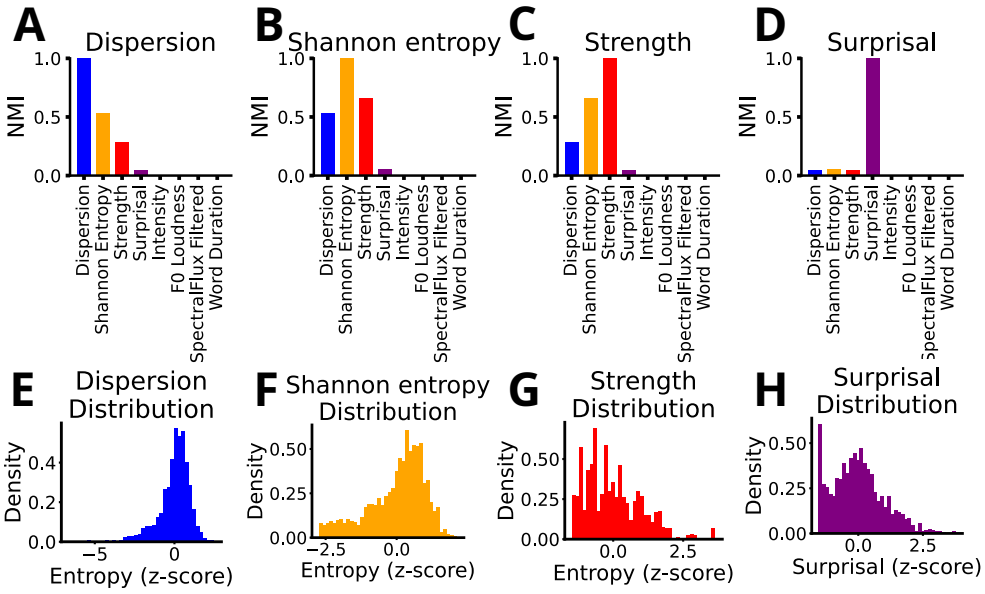


Fig. 7 Extended Data 1 (A-D) Normalised mutual information(NMI) between speech features and dispersion **(A)**, shannon entropy **(B)**, strength **(C)** and surprisal **(D)**. **(E-H)** Distribution of the values of disperson **(E)**, Shannon entropy **(F)**, strength **(G)** and Surprisal **(H)**

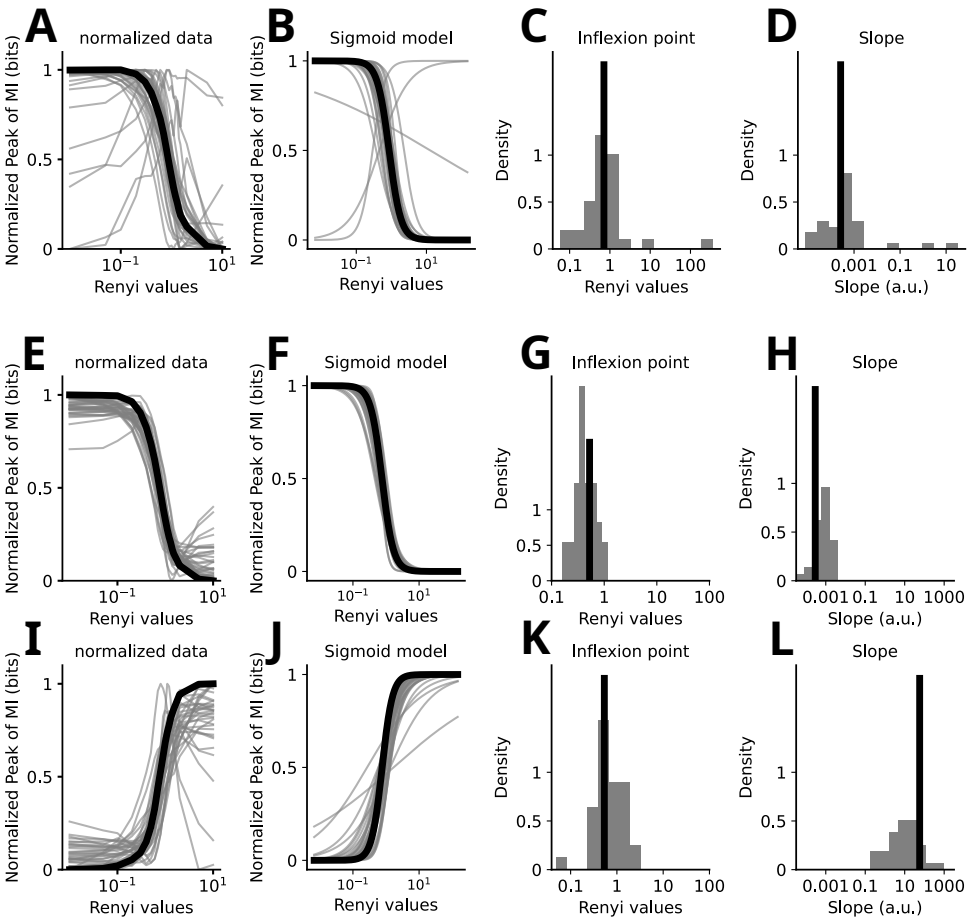


Fig. 8 Extended Data 2 (A-D) Individual entropy profiles for the average mutual information. Entropy profiles were normalized between 0 and 1 **(A)** and then a simple sigmoid was fitted onto the 20 data points per subjects. This resulted in a sigmoid model per subject **(B)**, from which a distribution of inflexion points **(C)** and slope **(D)** could be extracted. This procedure was repeated for each subject on the average mutual information in the dispersion cluster **(E-H) and in the strength cluster (I-L)**

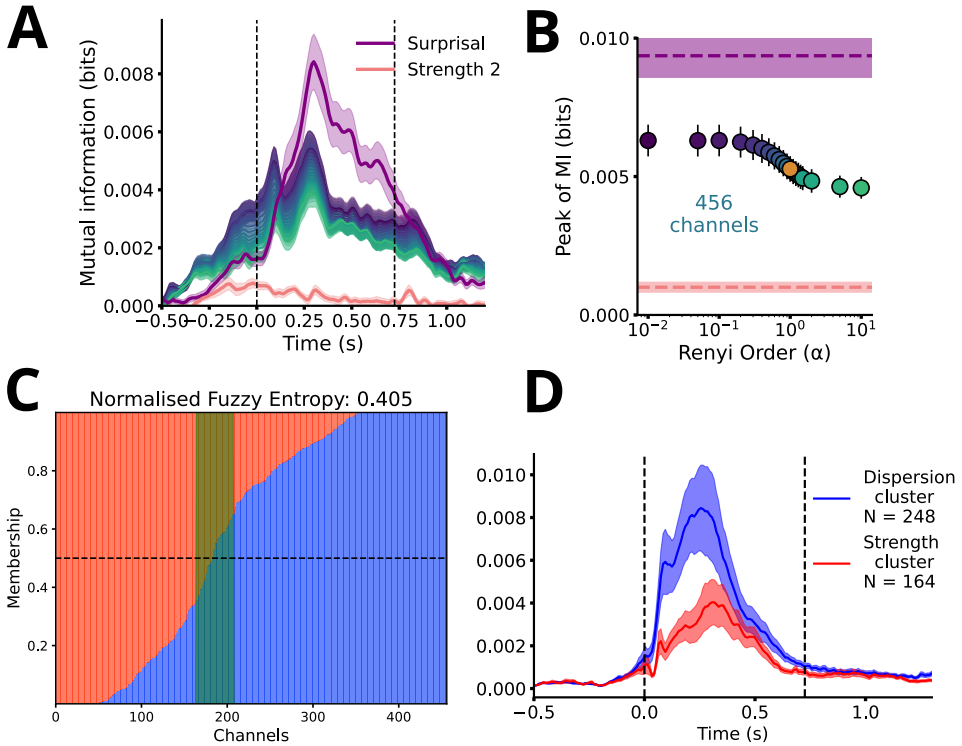


Fig. 9 Extended Data 3 (A,B) Grand average of MI over time across the 362 significant channels for different Renyi orders (α, color gradient), with a peak at 300 ms after word onset, on which the mutual information corresponding to surprisal(purple) and the entropy relative to the second most likely word(red) were represented. **(C)** Membership of the different channel relative to the dispersion(blue) and strength(red) clusters. in green, was represente the 0.65 threshold used in study. **(D)** Evoked reponse potential power relative to the onset of word for the dispersion and strength clusters.

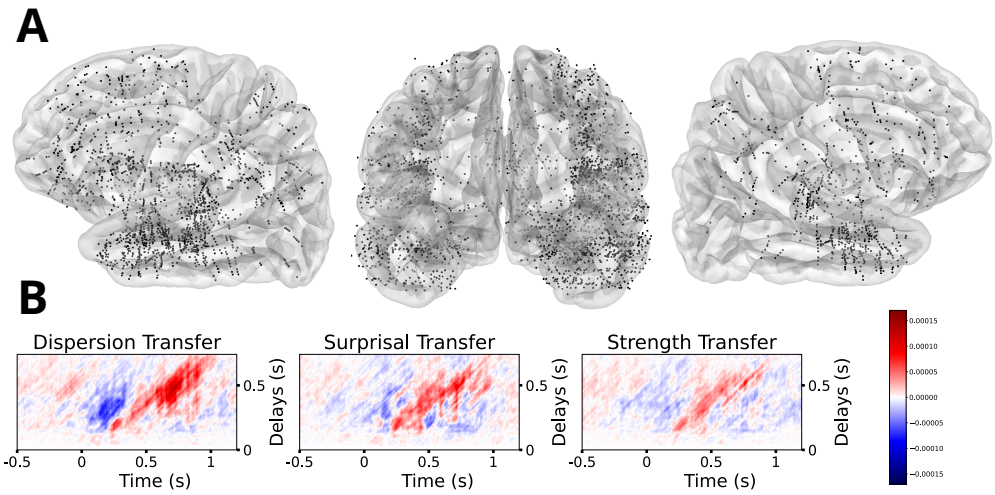


Fig. 10 Extended Data 4 (A) Overall implantation of all subjects pooled together. **(B)** Feature information transfer (FIT) time-delay matrix relative to dispersion, surprisal and strength.

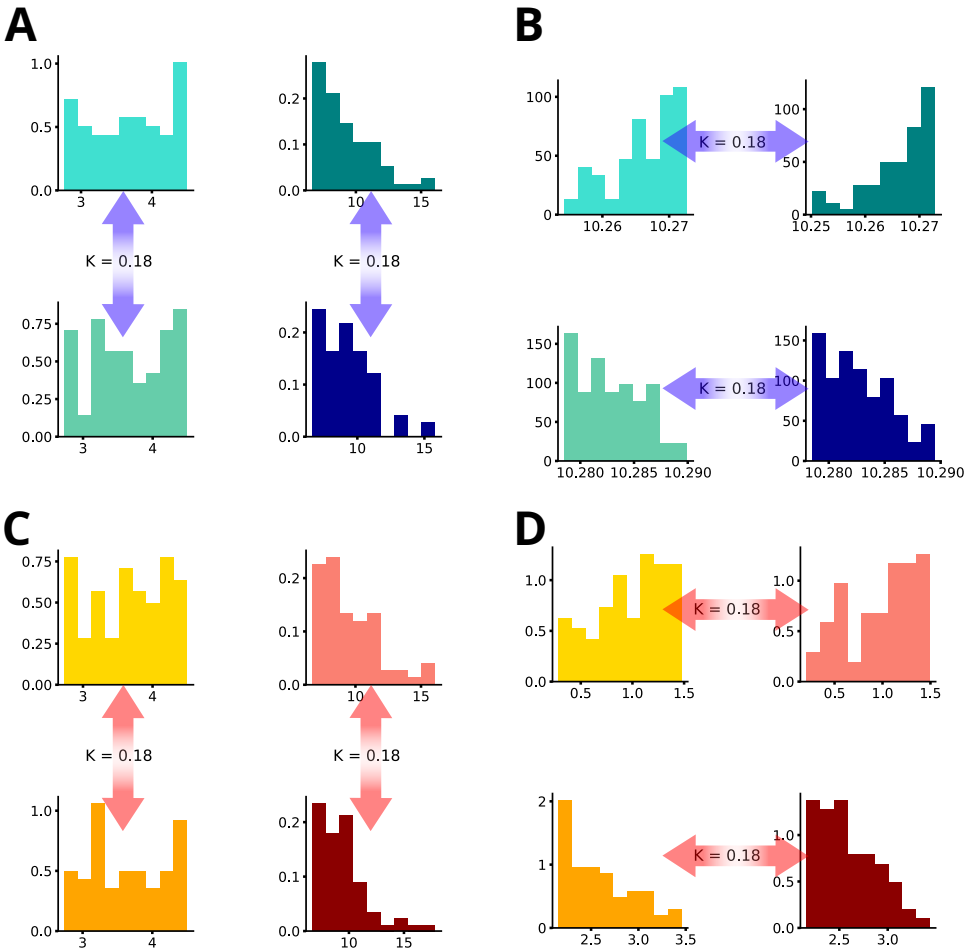


Fig. 11 Extended Data 5 (A) Distribution of dispersion across the dispersion-surprisal quartiles. Two quartiles vertically aligned present similar dispersion values (K = 0.18). **(B)** Distribution of surprisal across the dispersion-surprisal quartiles. Two quartiles horizontally aligned present similar surprisal values (K = 0.18). **(C)** Distribution of strength across the strength-surprisal quartiles. Two quartiles vertically aligned present similar strength values (K = 0.18). **(D)** Distribution of surprisal across the strength-surprisal quartiles. Two quartiles horizontally aligned present similar surprisal values (K = 0.18).

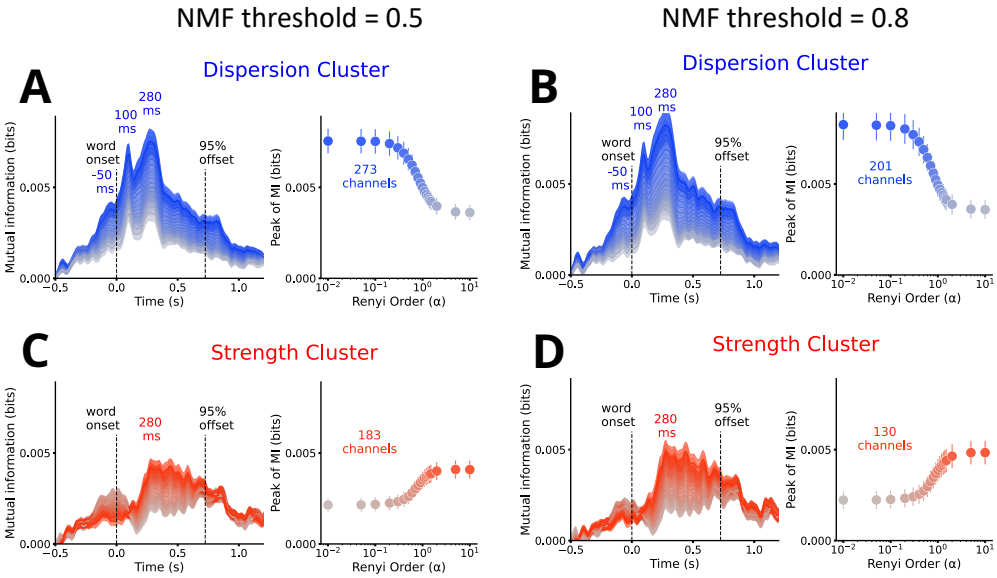


Fig. 12 Extended Data 6 (A-D) Grand average of MI over time across channels for different Renyi orders (α, color gradient) and MI peak as a function of Renyi order, averaged across channels. These were plotted for an NMF threshold of 0.5 for the dispersion **(A)** and strength **(C)** clusters, as well as for an NMF threshold of 0.8 for the dispersion **(B)** and strength **(D)** clusters.

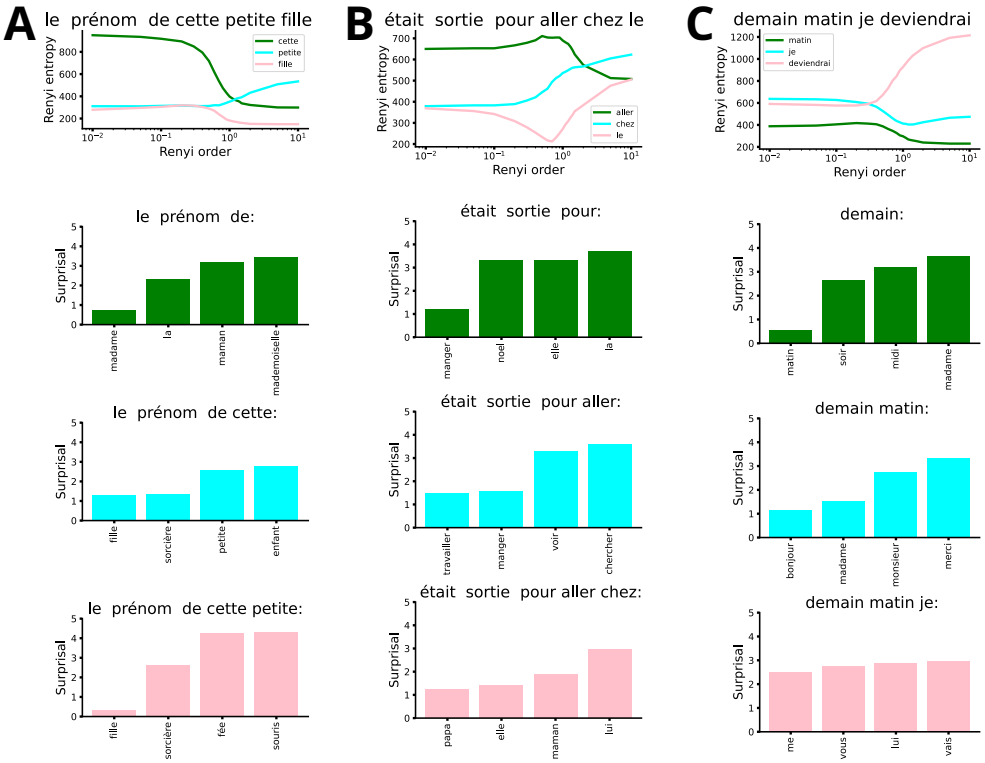
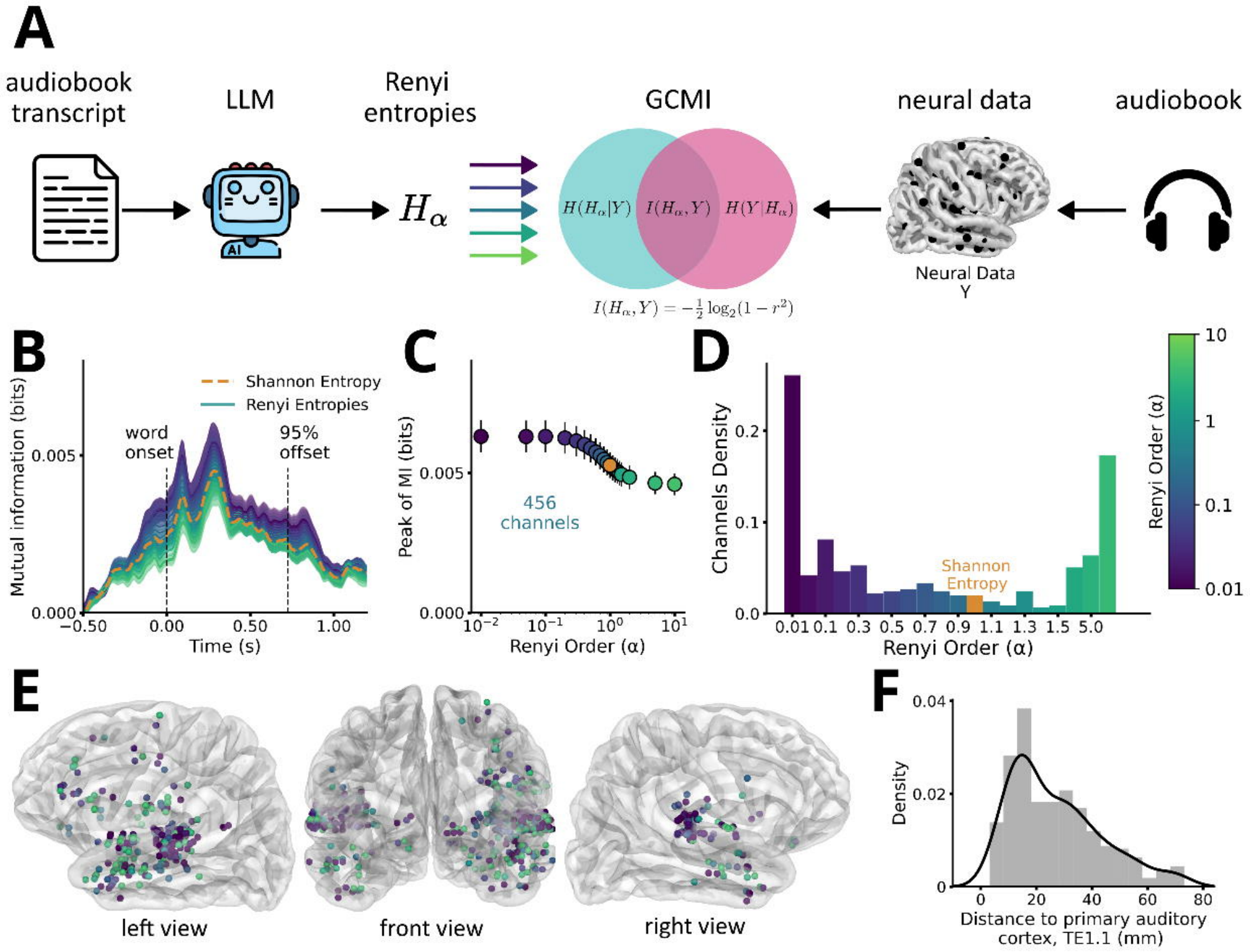
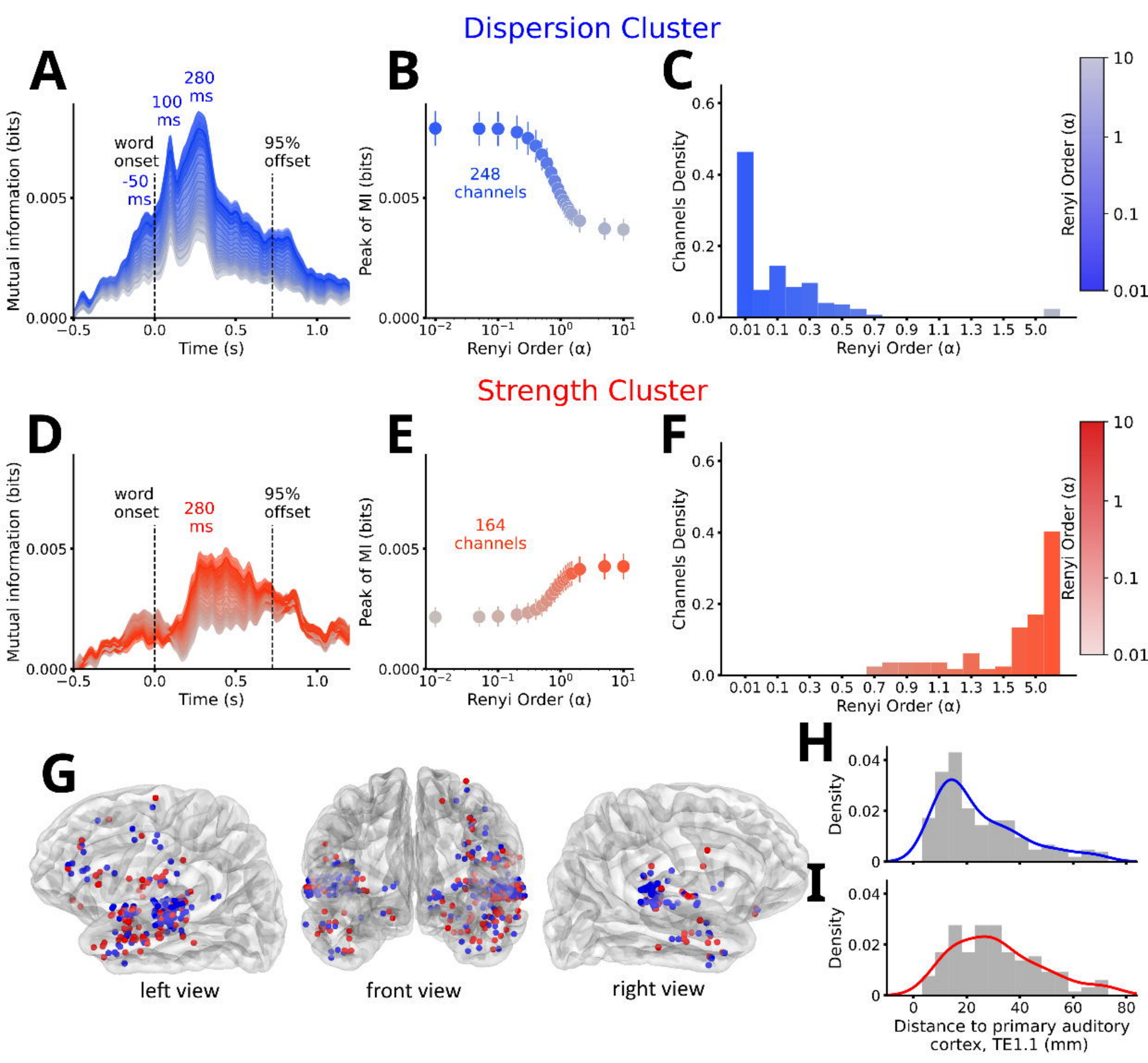
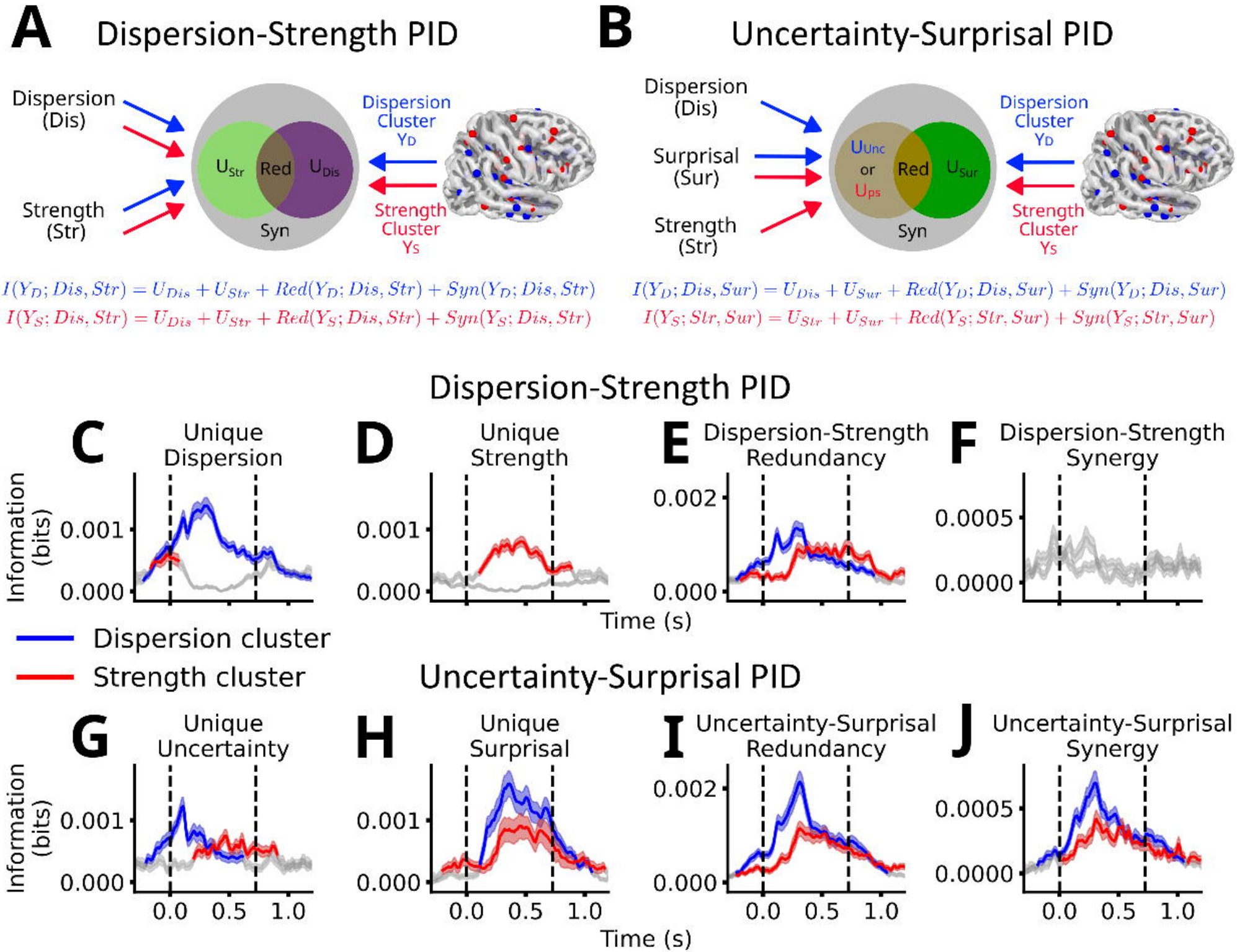
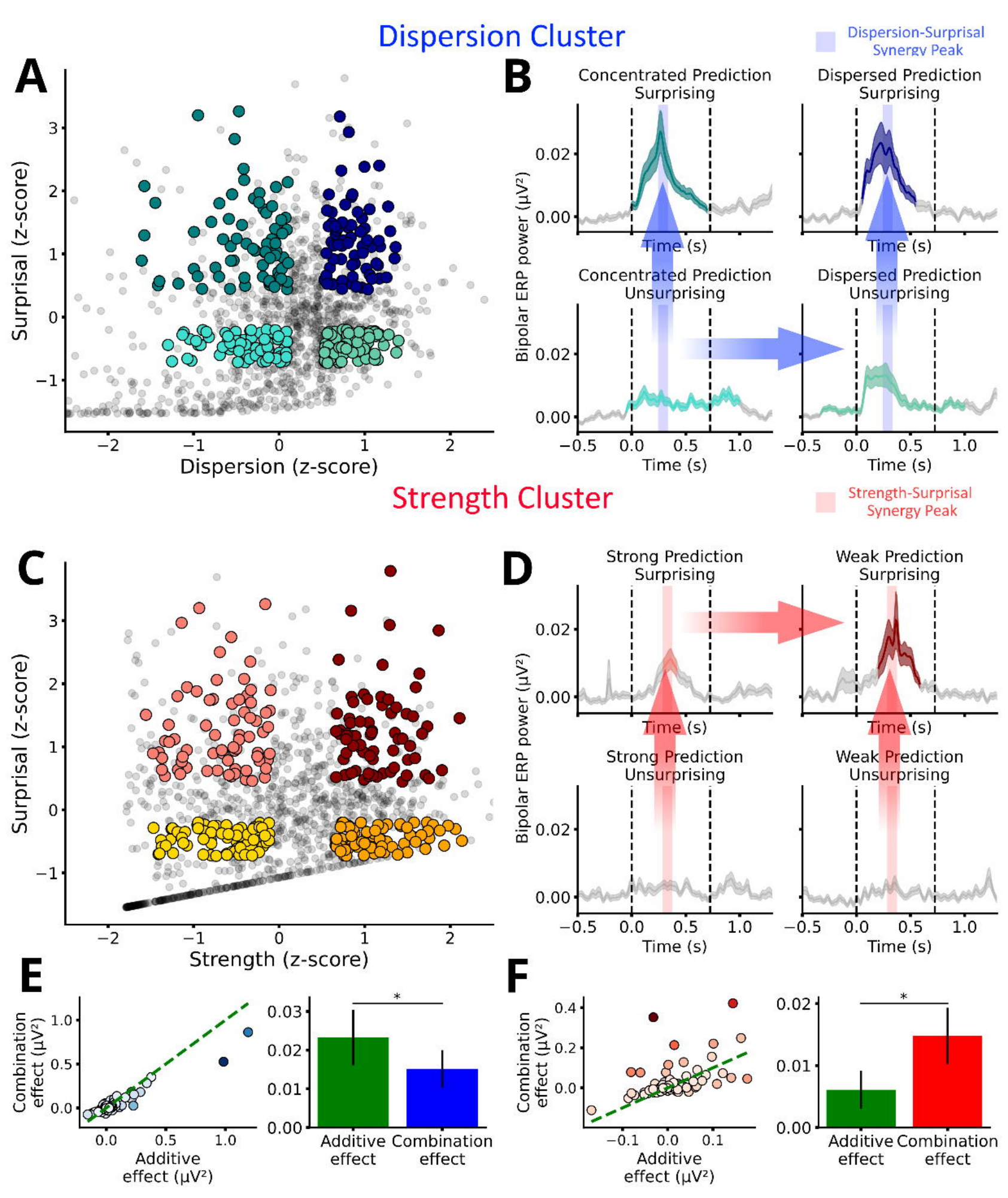


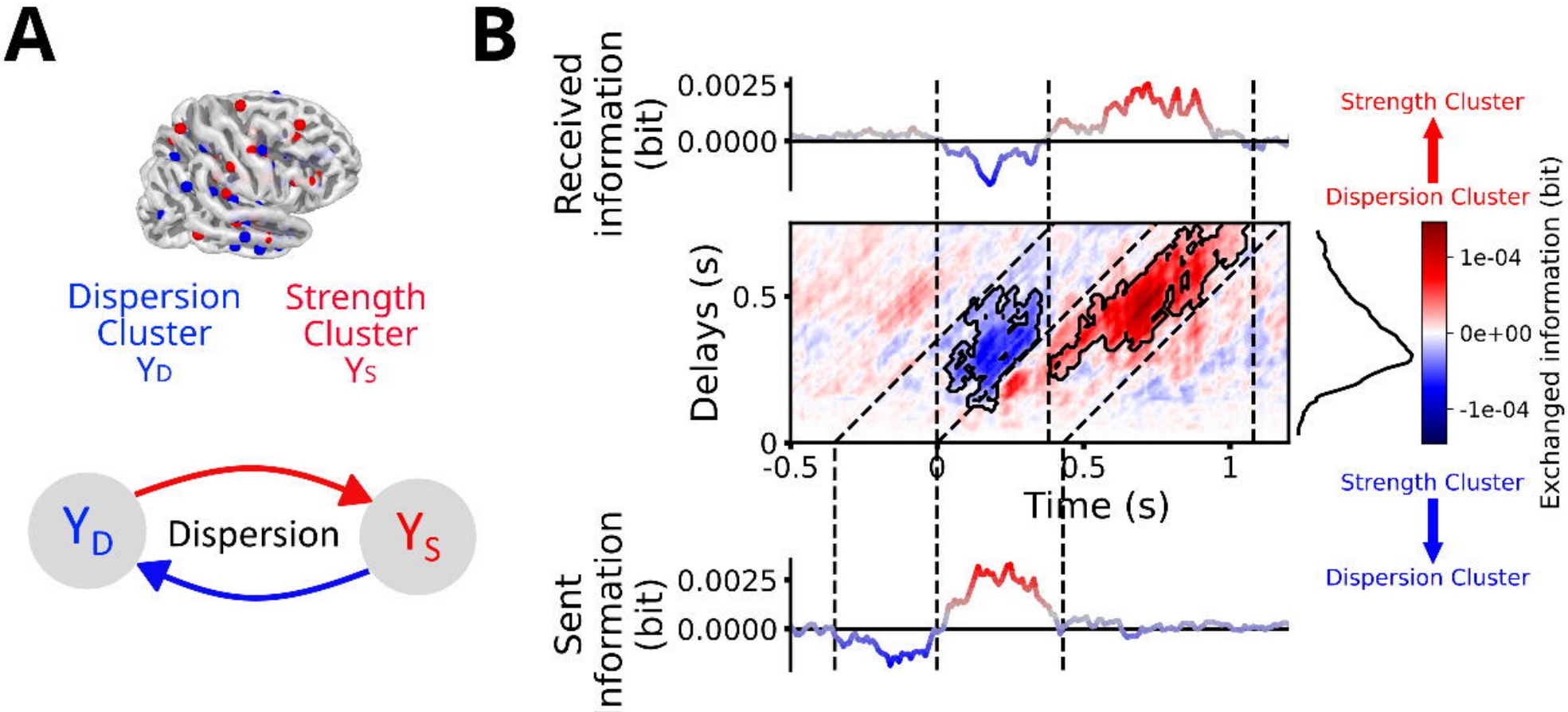
Fig. 13 Extended Data 7 (A-C) Example of sentences extracted from the dataset and the rank of their associated renyi entropy.

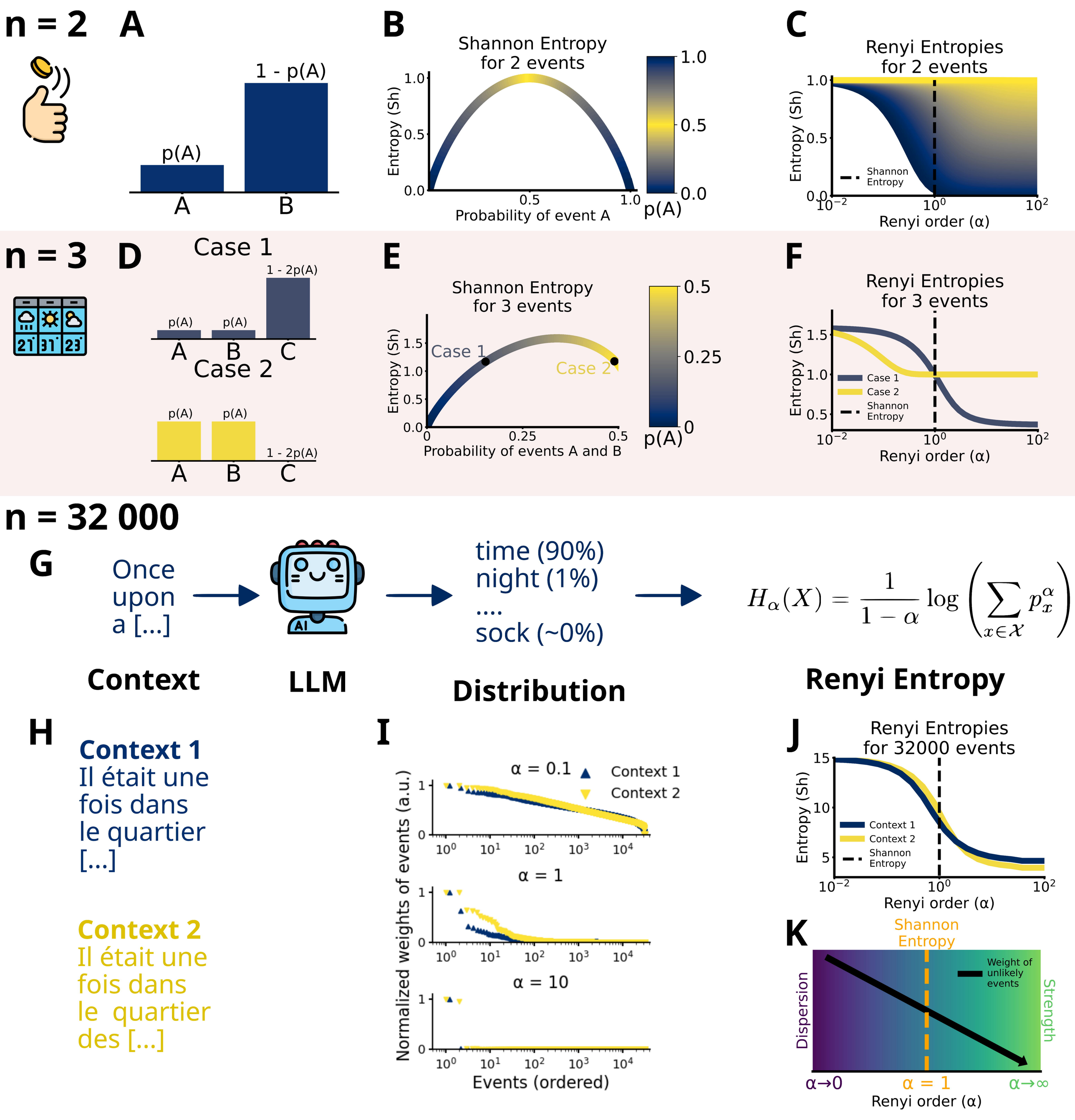

A
audiobook transcript
LLM
Renyi entropies
GCMI
neural data
audiobook
H_α
$H(H_\alpha|Y)$
$I(H_\alpha, Y)$
$H(Y|H_\alpha)$
Neural Data
Y
$I(H_\alpha, Y) = -\frac{1}{2}\log_2(1 - r^2)$
B
Shannon Entropy
Renyi Entropies
word onset
95% offset
Mutual information (bits)
0.005
0.000
−0.50
0.00
0.50
1.00
Time (s)
C
Peak of MI (bits)
0.005
0.000
456 channels
10^{-2}
10^{-1}
10^{0}
10^{1}
Renyi Order (α)
D
Channels Density
0.2
0.1
0.0
Shannon Entropy
0.01 0.1 0.3 0.5 0.7 0.9 1.1 1.3 1.5 5.0
Renyi Order (α)
Renyi Order (α)
10
1
0.1
0.01
E
left view
front view
right view
F
Density
0.04
0.02
0
0
20
40
60
80
Distance to primary auditory cortex, TE1.1 (mm)

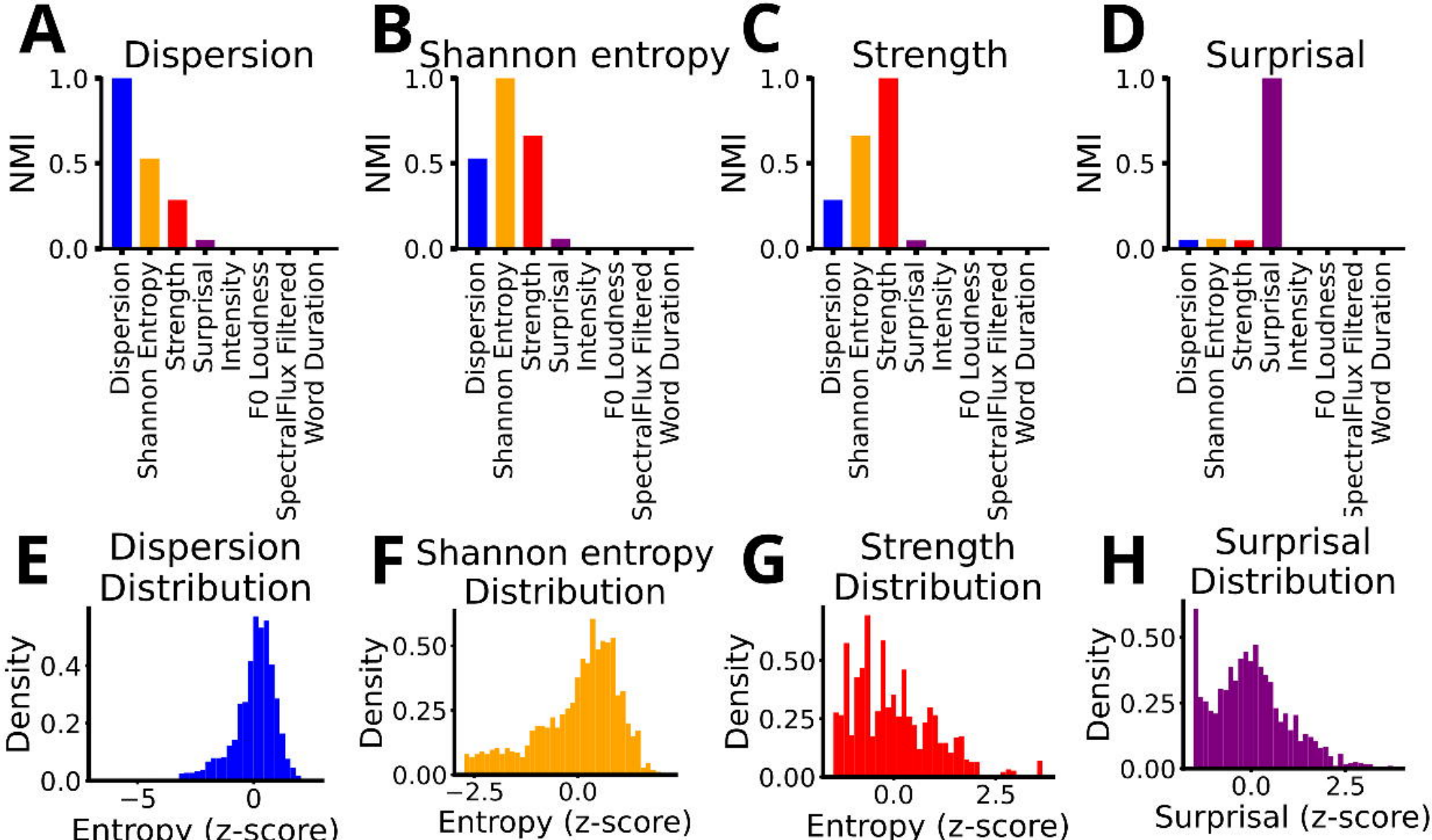

Dispersion Cluster
A
280 ms
100 ms
word onset
-50 ms
95% offset
Mutual information (bits)
Time (s)
B
248 channels
Peak of MI (bits)
Renyi Order (α)
C
Channels Density
Renyi Order (α)
Strength Cluster
D
word onset
280 ms
95% offset
Mutual information (bits)
Time (s)
E
164 channels
Peak of MI (bits)
Renyi Order (α)
F
Channels Density
Renyi Order (α)
G
left view
front view
right view
H
Density
I
Density
Distance to primary auditory cortex, TE1.1 (mm)

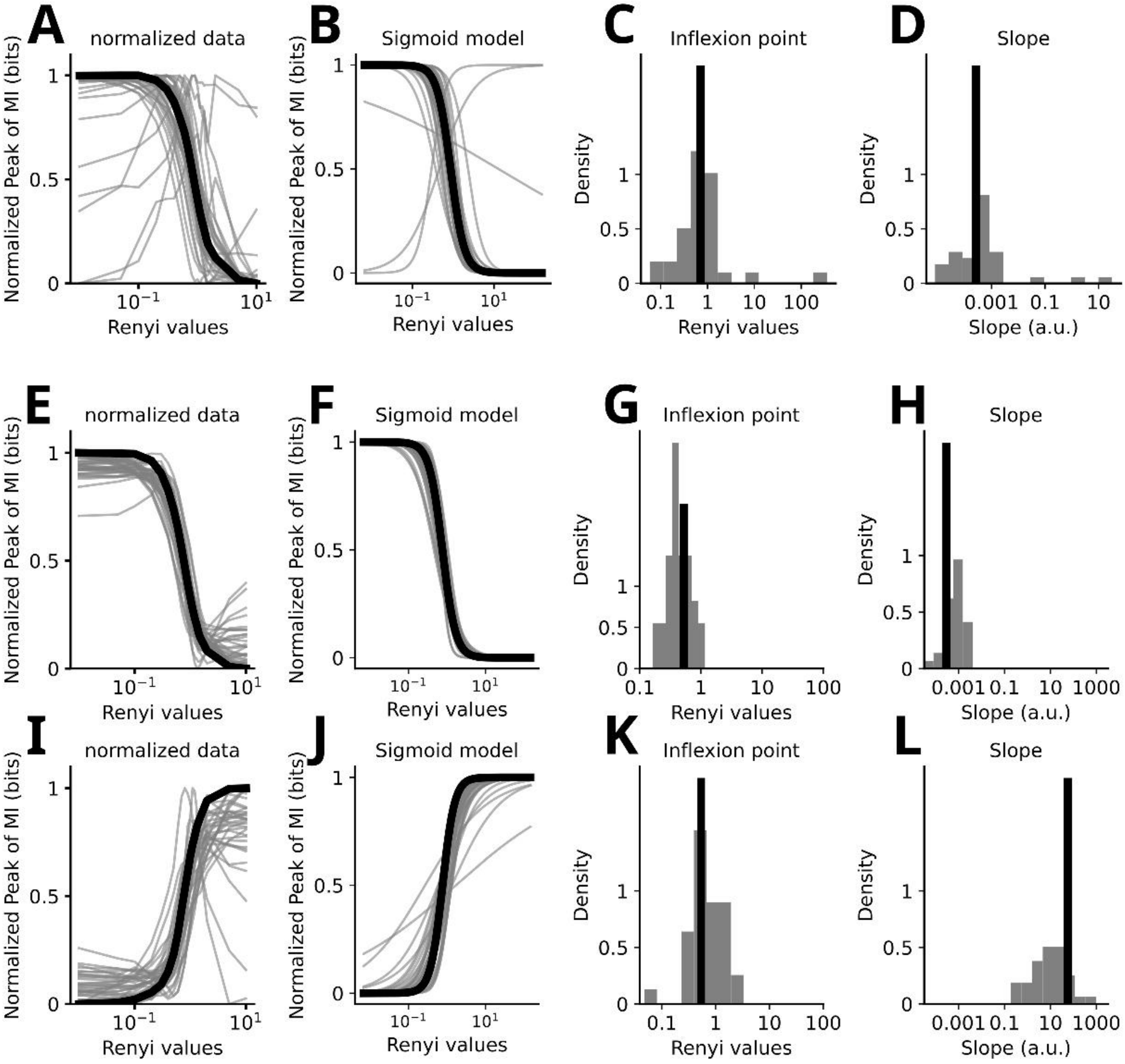

A
Dispersion-Strength PID
Dispersion (Dis)
Strength (Str)
U_Str
Red
U_Dis
Syn
Dispersion Cluster Y_D
Strength Cluster Y_S
B
Uncertainty-Surprisal PID
Dispersion (Dis)
Surprisal (Sur)
Strength (Str)
U_Unc or U_ps
Red
U_Sur
Syn
Dispersion Cluster Y_D
Strength Cluster Y_S
$I(Y_D; Dis, Str) = U_{Dis} + U_{Str} + Red(Y_D; Dis, Str) + Syn(Y_D; Dis, Str)$
$I(Y_S; Dis, Str) = U_{Dis} + U_{Str} + Red(Y_S; Dis, Str) + Syn(Y_S; Dis, Str)$
$I(Y_D; Dis, Sur) = U_{Dis} + U_{Sur} + Red(Y_D; Dis, Sur) + Syn(Y_D; Dis, Sur)$
$I(Y_S; Str, Sur) = U_{Str} + U_{Sur} + Red(Y_S; Str, Sur) + Syn(Y_S; Str, Sur)$
Dispersion-Strength PID
C Unique Dispersion
D Unique Strength
E Dispersion-Strength Redundancy
F Dispersion-Strength Synergy
Information (bits)
0.001
0.000
0.002
0.0005
0.0000
Time (s)
Dispersion cluster
Strength cluster
Uncertainty-Surprisal PID
G Unique Uncertainty
H Unique Surprisal
I Uncertainty-Surprisal Redundancy
J Uncertainty-Surprisal Synergy
0.0
0.5
1.0

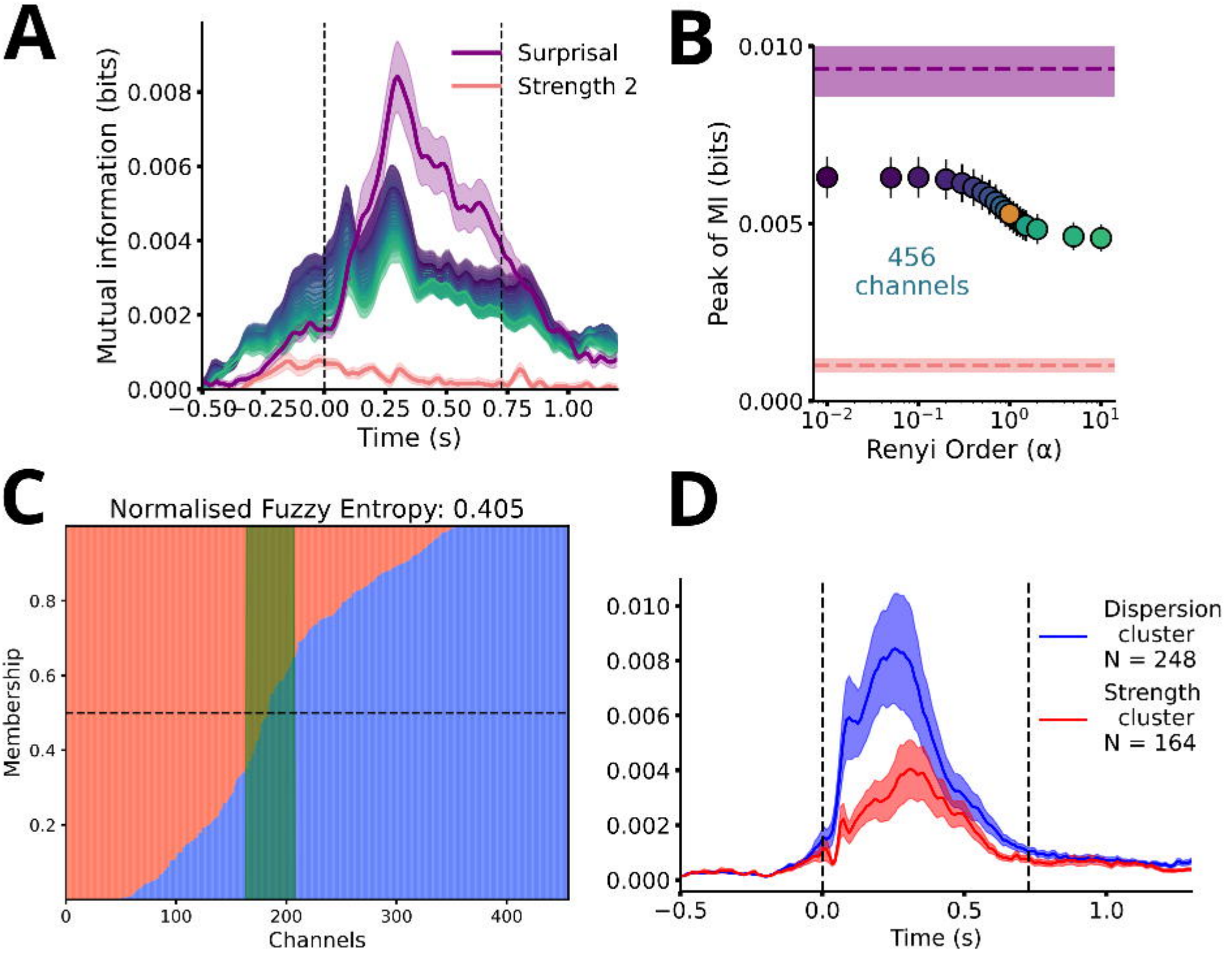

Dispersion Cluster
Dispersion-Surprisal Synergy Peak
A
Surprisal (z-score)
Dispersion (z-score)
B
Concentrated Prediction Surprising
Dispersed Prediction Surprising
Concentrated Prediction Unsurprising
Dispersed Prediction Unsurprising
Bipolar ERP power (μV²)
Time (s)
Strength Cluster
Strength-Surprisal Synergy Peak
C
Strength (z-score)
D
Strong Prediction Surprising
Weak Prediction Surprising
Strong Prediction Unsurprising
Weak Prediction Unsurprising
E
Combination effect (μV²)
Additive effect (μV²)
Additive effect
Combination effect
F

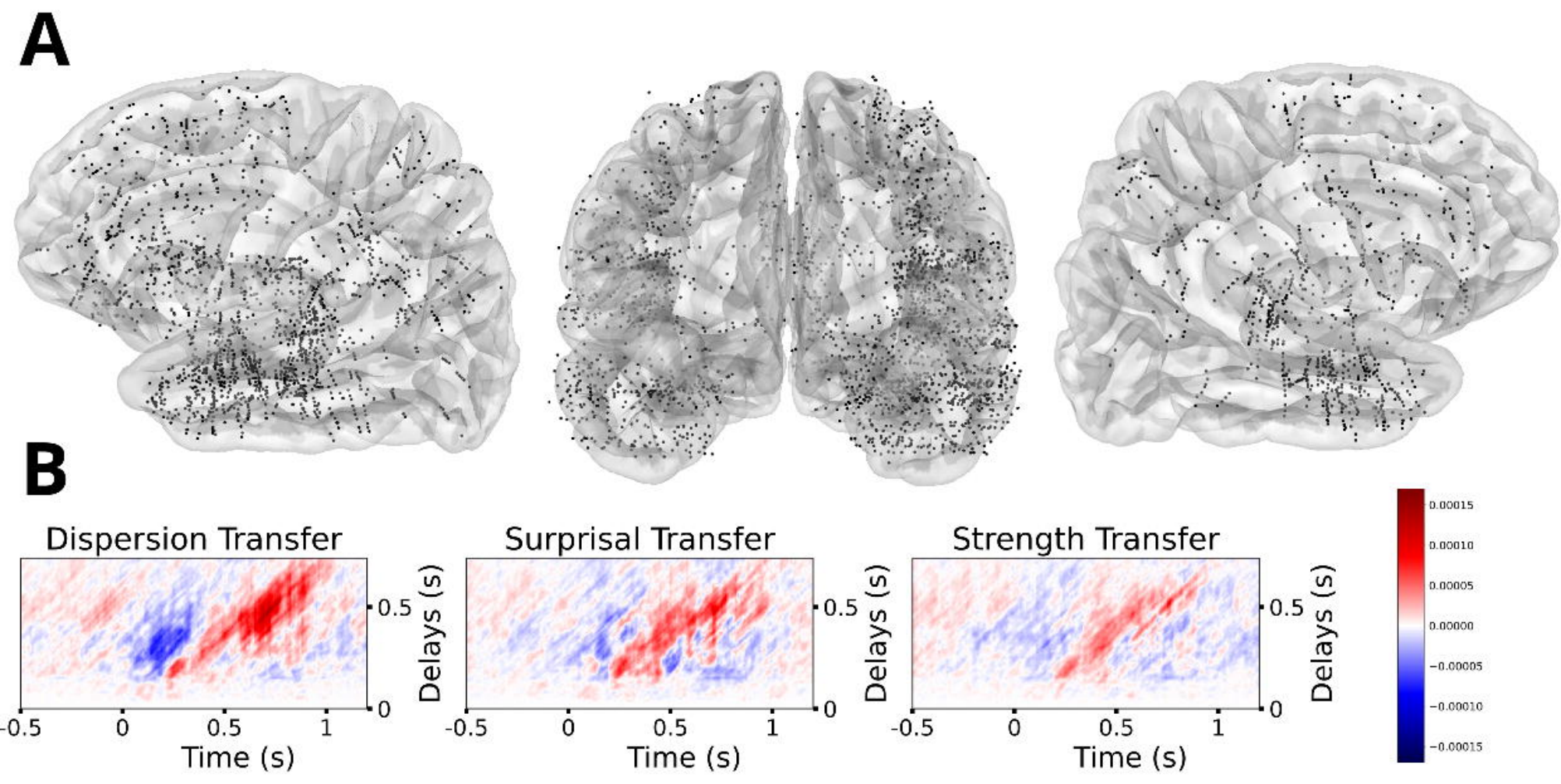

A
Dispersion Cluster
Y_D
Strength Cluster
Y_S
Y_D
Dispersion
Y_S
B
Received information (bit)
0.0025
0.0000
Delays (s)
0.5
0
-0.5
0
0.5
1
Time (s)
Sent Information (bit)
0.0025
0.0000
Strength Cluster
Dispersion Cluster
1e-04
0e+00
-1e-04
Exchanged information (bit)
Strength Cluster
Dispersion Cluster

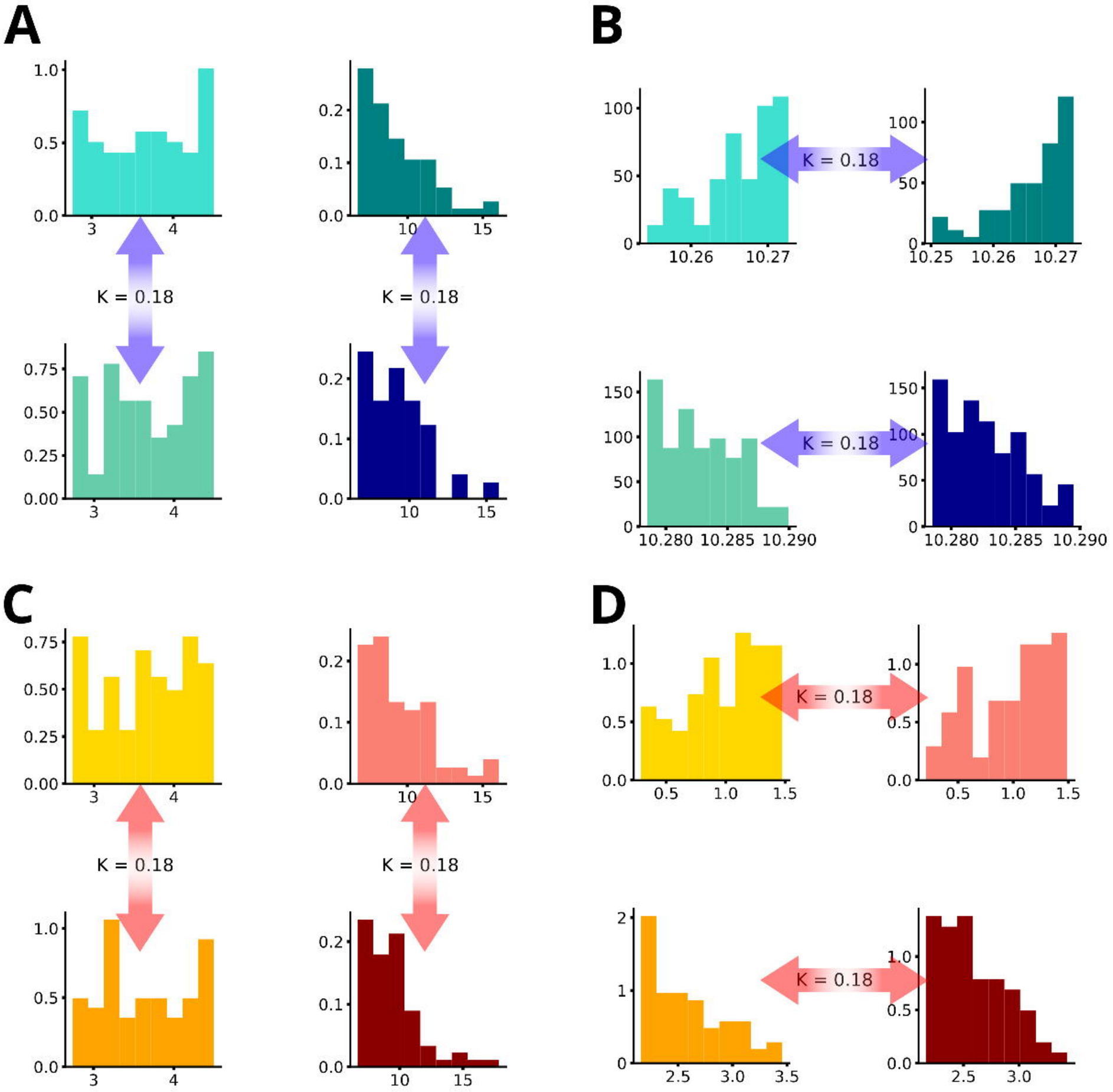

n = 2
A
1 - p(A)
p(A)
A
B
B
Shannon Entropy
for 2 events
Entropy (Sh)
Probability of event A
p(A)
C
Renyi Entropies
for 2 events
Entropy (Sh)
Shannon
Entropy
Renyi order (α)
n = 3
D
Case 1
1 - 2p(A)
p(A)
p(A)
A
B
C
Case 2
p(A)
p(A)
1 - 2p(A)
A
B
C
E
Shannon Entropy
for 3 events
Case 1
Case 2
Entropy (Sh)
Probability of events A and B
p(A)
F
Renyi Entropies
for 3 events
Entropy (Sh)
Case 1
Case 2
Shannon
Entropy
Renyi order (α)
n = 32 000
G
Once
upon
a [...]
AI
time (90%)
night (1%)
....
sock (~0%)
$H_\alpha(X) = \frac{1}{1-\alpha}\log\left(\sum_{x\in\mathcal{X}} p_x^\alpha\right)$
Context
LLM
Distribution
Renyi Entropy
H
Context 1
Il était une
fois dans
le quartier
[...]
Context 2
Il était une
fois dans
le quartier
des [...]
I
α = 0.1
Context 1
Context 2
α = 1
α = 10
Normalized weights of events (a.u.)
Events (ordered)
J
Renyi Entropies
for 32000 events
Entropy (Sh)
Context 1
Context 2
Shannon
Entropy
Renyi order (α)
K
Shannon
Entropy
Weight of
unlikely
events
Dispersion
Strength
α→0
α = 1
α→∞
Renyi order (α)

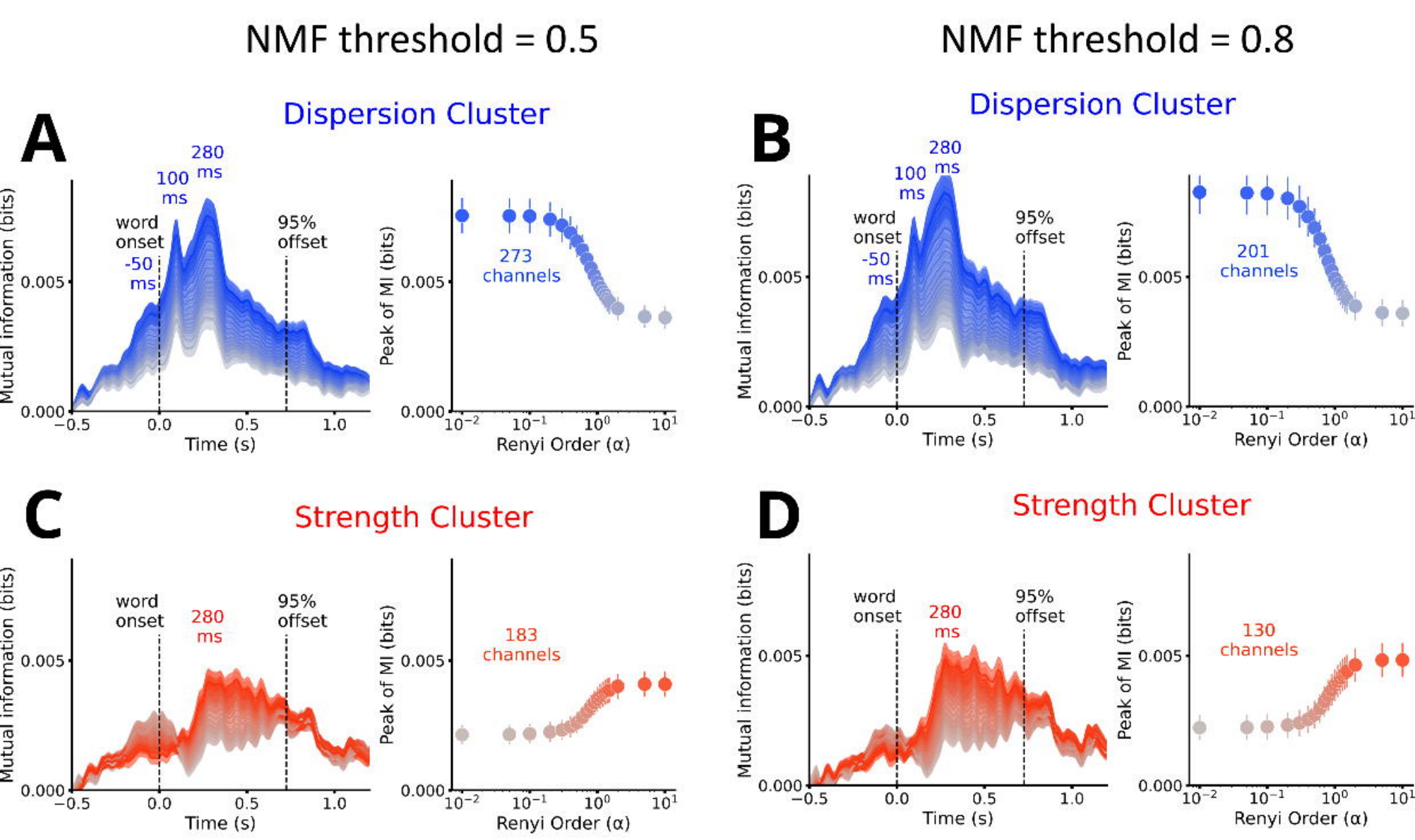

A
Dispersion
B
Shannon entropy
C
Strength
D
Surprisal
NMI
1.0
0.5
0.0
Dispersion
Shannon Entropy
Strength
Surprisal
Intensity
F0 Loudness
SpectralFlux Filtered
Word Duration
E
Dispersion Distribution
F
Shannon entropy Distribution
G
Strength Distribution
H
Surprisal Distribution
Density
0.4
0.2
0.0
−5
0
Entropy (z-score)
0.50
0.25
0.00
−2.5
0.0
2.5
Surprisal (z-score)

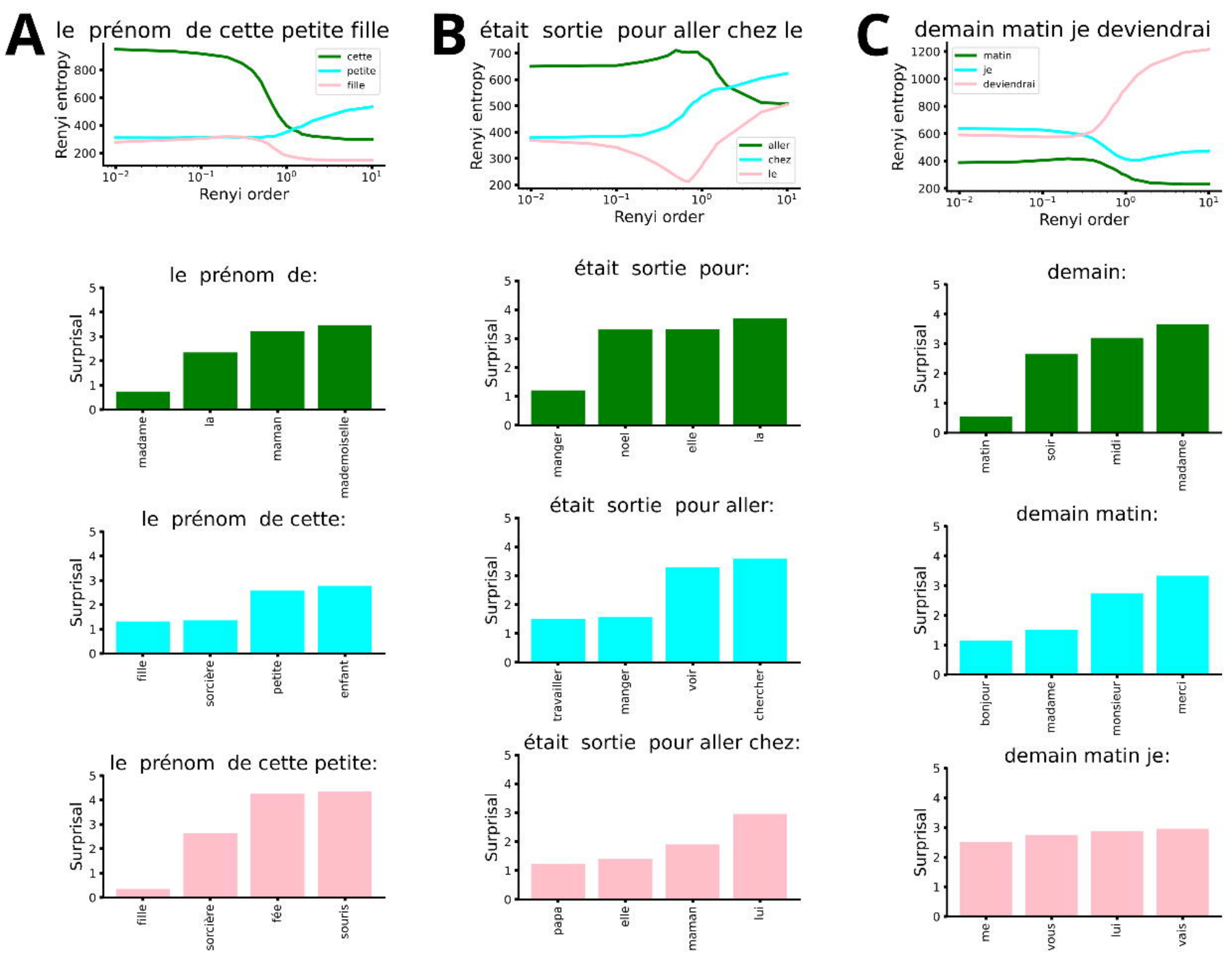

A
normalized data
Normalized Peak of MI (bits)
Renyi values
B
Sigmoid model
Normalized Peak of MI (bits)
Renyi values
C
Inflexion point
Density
Renyi values
D
Slope
Density
Slope (a.u.)
E
normalized data
Normalized Peak of MI (bits)
Renyi values
F
Sigmoid model
Normalized Peak of MI (bits)
Renyi values
G
Inflexion point
Density
Renyi values
H
Slope
Density
Slope (a.u.)
I
normalized data
Normalized Peak of MI (bits)
Renyi values
J
Sigmoid model
Normalized Peak of MI (bits)
Renyi values
K
Inflexion point
Density
Renyi values
L
Slope
Density
Slope (a.u.)


A
Mutual information (bits)
0.008
0.006
0.004
0.002
0.000
−0.50 −0.25 0.00 0.25 0.50 0.75 1.00
Time (s)
Surprisal
Strength 2
B
Peak of MI (bits)
0.010
0.005
0.000
456
channels
10^{-2} 10^{-1} 10^{0} 10^{1}
Renyi Order (α)
C
Normalised Fuzzy Entropy: 0.405
Membership
0.8
0.6
0.4
0.2
0 100 200 300 400
Channels
D
0.010
0.008
0.006
0.004
0.002
0.000
−0.5 0.0 0.5 1.0
Time (s)
Dispersion
cluster
N = 248
Strength
cluster
N = 164


A
B
Dispersion Transfer
Surprisal Transfer
Strength Transfer
Delays (s)
Time (s)
-0.5
0
0.5
1
0.5
0
0.00015
0.00010
0.00005
0.00000
−0.00005
−0.00010
−0.00015


A
K = 0.18
K = 0.18
B
K = 0.18
K = 0.18
C
K = 0.18
K = 0.18
D
K = 0.18
K = 0.18


NMF threshold = 0.5
NMF threshold = 0.8
A
Dispersion Cluster
280 ms
100 ms
word onset
-50 ms
95% offset
Mutual information (bits)
Time (s)
273 channels
Peak of MI (bits)
Renyi Order (α)
B
Dispersion Cluster
201 channels
C
Strength Cluster
183 channels
D
Strength Cluster
130 channels


A
le prénom de cette petite fille
Renyi entropy
Renyi order
cette
petite
fille
B
était sortie pour aller chez le
Renyi entropy
Renyi order
aller
chez
le
C
demain matin je deviendrai
Renyi entropy
Renyi order
matin
je
deviendrai
le prénom de:
madame
la
maman
mademoiselle
le prénom de cette:
fille
sorcière
petite
enfant
le prénom de cette petite:
fille
sorcière
fée
souris
était sortie pour:
manger
noel
elle
la
était sortie pour aller:
travailler
manger
voir
chercher
était sortie pour aller chez:
papa
elle
maman
lui
demain:
matin
soir
midi
madame
demain matin:
bonjour
madame
monsieur
merci
demain matin je:
me
vous
lui
vais
Surprisal