HAW HAMBURG

MASTER THESIS
Nick Alexander

Decoding Neural Representations of Expected and Actual Aversive Stimuli in Associative Learning Using fMRI-Based Multi-Voxel Pattern Analysis

Faculty of Engineering and Computer Science
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Nick Alexander

Decoding Neural Representations of Expected and Actual Aversive Stimuli in Associative Learning Using fMRI-Based Multi-Voxel Pattern Analysis

Master thesis submitted for examination in Master´s degree
in the study course *Master of Science Informatik*
at the Department Computer Science
at the Faculty of Engineering and Computer Science
at University of Applied Science Hamburg

Supervisor: Prof. Dr. Christian Lins
Supervisor: Dr. Lieven A. Schenk

Submitted on: 14 February 2025

Nick Alexander

Thema der Arbeit

Decoding Neural Representations of Expected and Actual Aversive Stimuli in Associative Learning Using fMRI-Based Multi-Voxel Pattern Analysis

Stichworte

Multi-Voxel Pattern Analysis, fMRT, assoziatives Lernen, Insularkortex, aversive Reize, schnelles Reversal-Lernen

Kurzzusammenfassung

Schnell wechselnde Lernparadigmen (Rapid Reversal Learning) stellen eine Herausforderung für das Verständnis dar, wie das Gehirn interne Repräsentationen erwarteter aversiver Reize bildet und aufrechterhält. Ein zuvor von Horing und Büchel (2022) [23] vorgestellter Datensatz, der zur Untersuchung von Vorhersagefehler-Signalen (Prediction Errors, PEs) in der Insula verwendet wurde, wurde erneut analysiert. In dieser Arbeit wurde untersucht, ob neuronale Muster, die mit konditionierten Erwartungen aversiver Stimuli–konkret schmerzhafte Hitze und laute Geräusche–assoziiert sind, mittels funktioneller Magnetresonanztomographie (fMRT) und Multi-Voxel Pattern Analysis (MVPA) robust dekodiert werden können. Siebenundvierzig Teilnehmende absolvierten ein Transreinforcer-Konditionierungsprotokoll mit häufigen, unangekündigten Wechseln der Hinweisreiz-Outcome-Kontingenzen, wodurch eine längerfristige Stabilisierung der Erwartungen verhindert wurde. Im Mittelpunkt der Untersuchung stand der insuläre Kortex, eine Region, die sowohl an der Antizipation als auch an der Verarbeitung aversiver Stimuli beteiligt ist. Nach einer gründlichen Kontrolle zeitlicher und sitzungsbezogener Störfaktoren mittels systematischer Anwendung des Same Analysis Approach (SAA) [19] und einer auf ganzzahliger linearer Programmierung basierenden Kreuzvalidierungs-Optimierung–was eine Reduktion der Analyse von 128 auf 32 Durchgänge pro Teilnehmenden erforderlich machte–zeigte sich eine konsistent überzufällige Dekodierbarkeit nur für die tatsächliche Stimulusmodalität (also ob Teilnehmende schmerzhafte Hitze oder laute Geräusche erhielten). Im Gegensatz dazu ließ sich die erwartete Stimulusmodalität (das heißt, was die Teilnehmenden zu erhalten glaubten) unter den Bedingungen schneller Kontingenswechsel nicht zuverlässig aus der insulären Aktivität dekodieren. Dies könnte entweder auf die schnellen Wechsel oder auf bislang nicht identifizierte Störfaktoren

zurückzuführen sein, die feine, erwartungsbezogene Muster maskieren. Whole-Brain-Searchlight-Analysen bestätigten diese Befunde durch den Nachweis verteilter Cluster überzufälliger Dekodierung für die tatsächlichen Stimulusmodalitäten, während keine konsistenten Signaturen der erwarteten Modalität erkennbar waren. Die Prävalenzanalyse [1] zeigte, dass die überzufällige Dekodierung der tatsächlichen Modalität bei der Mehrheit der Teilnehmenden vorlag, während es keine deutlichen Hinweise darauf gab, dass sich die Dekodierung der erwarteten Modalität über die gesamte Stichprobe generalisieren ließ. Diese Ergebnisse deuten darauf hin, dass stabile Konditionierungsphasen und Analyseansätze, die es erlauben, mehr Durchgänge zu nutzen und gleichzeitig Störfaktoren zu kontrollieren, für die Untersuchung neuronaler Repräsentationen aversiver Erwartungen vorteilhaft sein könnten. Zukünftige Untersuchungen sollten weiterhin potenzielle Störfaktoren erforschen, die subtile, erwartungsbezogene Muster maskieren könnten.

Nick Alexander

Title of Thesis

Decoding Neural Representations of Expected and Actual Aversive in Associative Learning Using fMRI-Based Multi-Voxel Pattern Analysis

Keywords

multi-voxel pattern analysis, fMRI, associative learning, insular cortex, aversive stimuli, rapid reversal learning

Abstract

Rapid reversal learning paradigms pose a challenge for understanding how the brain forms and maintains internal representations of expected aversive outcomes. A dataset previously introduced by Horing and Büchel (2022) [23] to study insular prediction error (PE) signals was leveraged to determine whether neural patterns associated with conditioned expectations of aversive stimuli—specifically painful heat and loud sound—could be robustly decoded using functional magnetic resonance imaging (fMRI) and Multi-Voxel Pattern Analysis (MVPA). Forty-seven participants completed a transreinforcer

conditioning protocol with frequent, unannounced reversals of cue–outcome contingencies, preventing prolonged stabilization of expectations. The insular cortex, a region implicated in both the anticipation and the experience of aversive stimuli, was the primary focus. After thorough control for temporal and session-related confounds via systematic application of the Same Analysis Approach (SAA) [19] and Integer Linear Programming (ILP)-based cross-validation (CV) optimization, which necessitated reducing the analysis from 128 to 32 trials per subject, consistent above-chance decoding emerged only for actual stimulus modality (i.e., whether participants received painful heat or loud sound). By contrast, expected stimulus modality (i.e., what participants believed they would receive) could not be reliably decoded from insular activity under the rapid reversal conditions, though this could reflect either the rapid reversals or unidentified confounds masking subtle expectation-related patterns. Whole-brain searchlight analyses corroborated these findings, revealing distributed clusters of above-chance decoding for actual stimuli but failing to identify consistent signatures of expected modality. Prevalence inference [1] indicated that the above-chance decoding of actual modality was present in the majority of participants, whereas there was no evidence that expected modality decoding generalized broadly across the sample. These findings suggest that stable conditioning phases and analysis approaches that can utilize more trials while controlling for confounds may be beneficial for studying neural representations of aversive expectations. Future work should continue to investigate potential confounding factors that might mask subtle expectation-related patterns.

Contents

List of Figures

List of Tables

1 Introduction

A longstanding question in cognitive and affective neuroscience[1] concerns how the brain acquires and maintains internal representations of future aversive events. Classical conditioning paradigms have shown that organisms learn not only from the direct experience of an aversive unconditioned stimulus (US)—such as painful heat or loud sound—but also from a conditioned stimulus (CS) that predicts these outcomes [40]. Theoretical models generally agree that these learned expectations guide behavior and can profoundly shape both subjective experience and physiological responses [4, 38]. However, whether and how the brain encodes these CS-based aversive predictions remains an active area of investigation [3, 24].

1.1 Decoding Aversive Expectations with MVPA

In parallel with conceptual advances in learning theory, technical developments in neuroimaging have expanded the ability to detect distributed neural representations [20, 36]. Traditional univariate analyses often treat each brain voxel independently, focusing on regional signal intensity changes in relation to specific events. MVPA, by contrast, can reveal subtle and spatially distributed activity patterns encoding states such as reward, fear, or pain [26]. Particularly in the domain of aversive learning, MVPA methods have shown that it is possible to decode whether a participant is experiencing a painful or non-painful stimulus [47] and, in some cases, distinguish between different forms of aversive or negative outcomes [27]. While many studies have emphasized the actual delivery of aversive stimuli [6, 27, 30, 47], recent work has demonstrated that MVPA can also decode neural representations of anticipated pain, though these patterns appear distinct from those associated with actual pain experience [9]. However, it remains unclear whether

[1] Cognitive neuroscience studies brain mechanisms underlying mental processes like memory and decision-making, while affective neuroscience focuses specifically on emotions and emotional responses.

CS-based expectations can be reliably decoded under conditions of rapid learning and frequent contingency reversals.

1.2 The Challenge of Decoding Conditioned Expectations

Decoding whether someone expects a particular aversive stimulus (e.g., pain) is considerably more challenging than determining which actual stimulus they receive. This difficulty arises because "top–down" mental processes—such as attention levels or beliefs about what will happen—can modulate brain signals in regions that process pain or other unpleasant sensations. These modulations present an opportunity to detect "expectation signals" in those brain regions [9, 24].

However, any brain activity related to "I believe this is going to happen" can blend with broader influences, like how alert or anxious someone feels in general, potentially masking the expectation signal [43]. Traditional univariate analyses have revealed consistent brain responses during pain anticipation under controlled experimental conditions [37]. Yet, decoding specific expected outcomes from neural patterns can be especially challenging in environments with frequently shifting cue–outcome contingencies. Under these circumstances, participants must rapidly update their predictions, leaving limited time for robust, decodable neural representations of expectation to emerge. As a result, the constantly evolving or short-lived nature of these representations may hinder reliable decoding from fMRI data.

1.3 The Insular Cortex as a Prime ROI

A key region of interest (ROI) for decoding these aversive expectations is the insular cortex—particularly the anterior insula (AI). Across numerous studies, the insula has shown reliable activation not only during the experience of pain or other aversive stimuli [12] but also during anticipation or prediction of such events [9, 24]. Several properties make the insula especially relevant for the aims:

- **Aversive Processing and Prediction:** The AI consistently encodes both the intensity of painful stimulation and the mismatch between expected and actual outcomes [14, 23]. Furthermore, it appears involved in representing the salience of

aversive stimuli—i.e., how inherently attention-grabbing or important a stimulus is—responding robustly to highly salient or novel events across modalities [32, 41].

- **Multimodal Integration:** The AI receives inputs spanning different sensory modalities and has extensive connections to limbic and autonomic regions[2] [12]. This connectivity facilitates integration of predictive cues with potential outcomes—even when they switch between, for instance, painful heat and loud sound. Supporting this, [9] found that the anterior insula harbors shared representations between anticipated pain and other aversive experiences, while maintaining distinct patterns for expected versus actual pain.
- **Sensitivity to Conditioning:** Studies indicate that the insula's activity patterns can distinguish different anticipated aversive outcomes. For instance, [24] documented distinct anterior and posterior insula activation when participants learned to discriminate between interoceptive and exteroceptive threats[3].

Given these converging roles in aversive anticipation, salience, and prediction, the insula represents a strong candidate region for identifying whether neural patterns differentiate expected aversive modalities before the US occurs.

1.4 Frequent Reversals: Implications for Decoding

Several neuroimaging experiments seeking to study the flexibility of learning and prediction employ reversal paradigms, wherein associations between CS and US swap unpredictably [8]. Although these designs elucidate adaptive updating of predictions, they may hinder the formation of robust, decodable representations of particular aversive expectations [5]. Rapid or frequent reversals can mean that, just as an internal model of "CS $\rightarrow$ painful heat" has begun to stabilize, a switch occurs to "CS $\rightarrow$ loud sound." This volatility could yield null or inconsistent results when decoding expected modality, because the relevant neural patterns might not stabilize before the contingency shifts again. Indeed, frequent reversals can inadvertently introduce confounding correlations between trial order and experimental conditions, leading to false positives or below-chance classification if not carefully counterbalanced [19, 43, 46].

[2] The limbic and autonomic systems are neural networks controlling emotion/behavior and involuntary bodily functions (like heart rate and digestion) respectively.

[3] Interoceptive and exteroceptive refer to sensory stimuli originating from within the body (e.g., hunger, pain, temperature) and outside the body (e.g., sight, sound, touch) respectively.

Importantly, the present dataset—originally collected and analyzed by Horing and Büchel (2022) [23] for prediction-error research—was optimized for detecting ongoing updates in participants' predictions rather than expectation representations. Consequently, frequent reversals were designed to trigger ongoing updates in participants' predictions, but this limits the viability of straightforward decoding approaches aimed at anticipatory signals. As a result, extensive post hoc confound control—such as balancing trial order effects via integer linear programming (ILP) and validating label distributions using the Same Analysis Approach—is needed. Otherwise, one risks conflating genuine expectation-related signals with artifacts driven by trial progression or session drift.

1.5 Motivation and Outline

This work aims to determine whether CS expectations—specifically "which aversive modality do I believe is next?"—can be decoded from fMRI data. Given the established significance of the insula for aversive anticipation and salience processing, analyses were concentrated on this ROI. The dataset used features a transreinforcer conditioning paradigm[4], in which participants occasionally experience frequent, unannounced CS–US reversals that may complicate decodable expectation formation.

Because this paradigm was primarily intended to study rapid shifts in prediction error signals, the design is not inherently optimized to capture expectation representations. In turn, two key methodological strategies were adopted to mitigate the effects of this design limitation:

1. **Systematic Confound Control:** The Same Analysis Approach [19] is employed to detect whether variables like trial number or session number could be inadvertently decoded under standard CV. By iteratively testing the decoding pipeline on these confounders and then optimizing CV folds via ILP, spurious effects from trial order or session identity are minimized [19, 46].

2. **Prevalence Inference:** Beyond classic second-level tests of mean decoding accuracy, a prevalence-based statistical framework [1] is applied to determine how many participants exhibit above-chance decoding. This approach clarifies whether

[4] A transreinforcer paradigm uses different types of reinforcement or outcomes (in this case switching between pain and sound) within the same experimental setup.

any observed decoding is broadly typical of the population or limited to a few individuals.

Ultimately, the findings will indicate whether the insular cortex can encode distinct expected aversive modalities—painful heat versus loud sound—even under conditions of frequent reversals. If strong decoding is absent or inconsistent, this would reinforce the notion that learned expectations require longer consolidation periods or reduced volatility to be detectable in neural signals. Conversely, evidence of robust decoding would suggest that, despite high contingency volatility, neural representations of "what is expected to happen" can be captured through MVPA.

2 Methods

A comprehensive fMRI data analysis pipeline, as depicted in Figure 2.1, was implemented to ensure robust and reliable results.

The initial stages of this pipeline were conducted by [23]: the **Experimental Design** phase (Section 2.1), which follows established protocols for participant recruitment, trial structure, and stimulus presentation to elicit robust conditioned expectations.

In the **Data Acquisition and Preprocessing** stage (Sections 2.2 and 2.3), [23] performed the core preprocessing steps, including slice timing correction, motion correction, co-registration, and Least Squares Separate (LSS)-based neural response extraction [33]. The present work extends this stage by implementing ROI transformation and thresholding procedures.

Starting with the **Confound Control and Cross-Validation Optimization** stage (Sections 2.6 and 2.7) begins the primary novel contribution of this work. This stage implements careful trial selection criteria and introduces an ILP method to optimize CV design. This optimization ensures that potentially confounding variables such as trial order and session effects are balanced across folds, thereby enhancing the validity of subsequent analyses.

Finally, the **Multi-Voxel Pattern Analysis** stage (Section 2.8) employs two complementary approaches: ROI-based decoding with group-level prevalence inference, and whole-brain searchlight mapping. This dual approach provides both targeted analysis of anatomically defined regions and a comprehensive view of information distribution across the brain, enabling robust characterization of neural patterns associated with both expected and actual stimulus modalities.

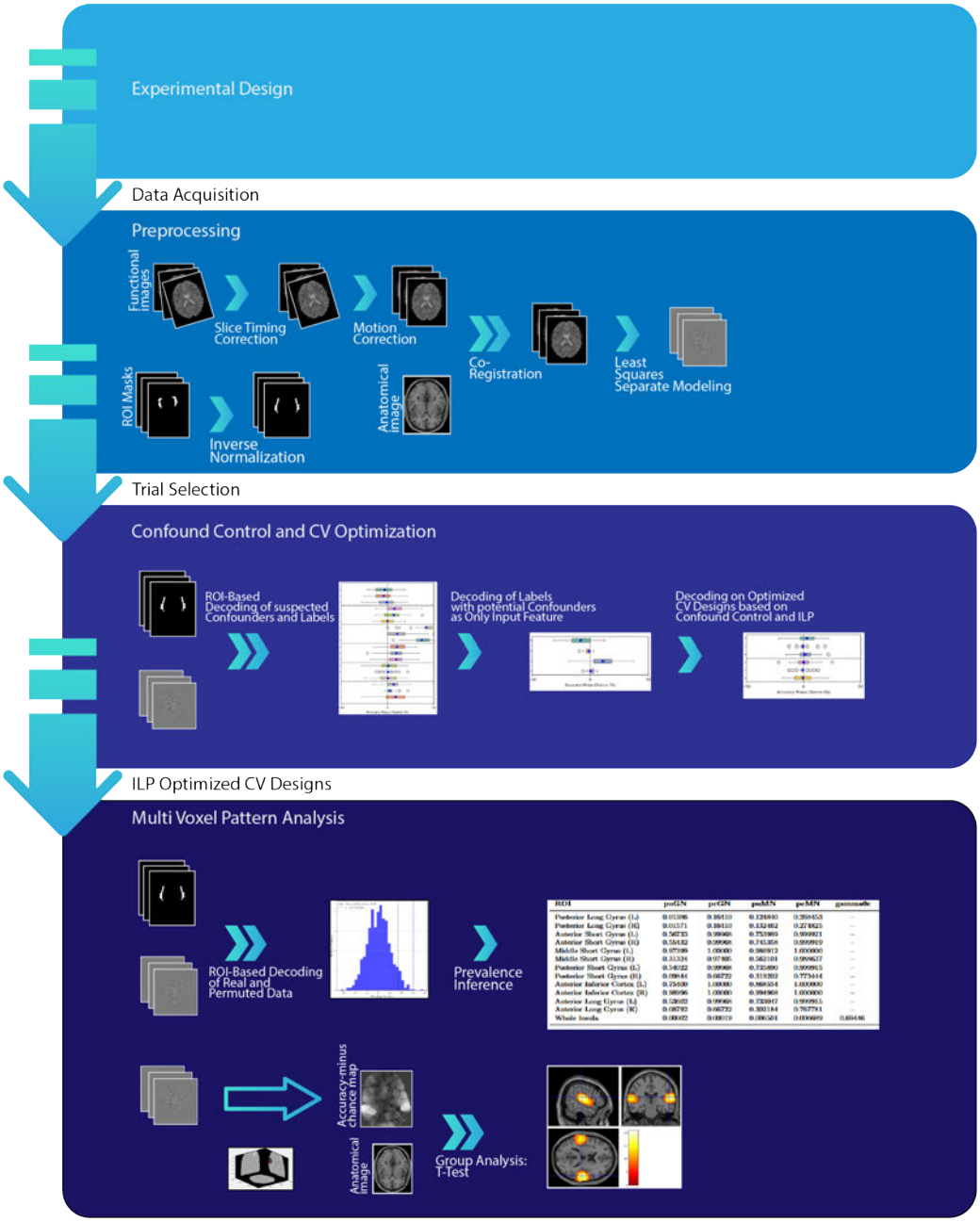

Figure 2.1: **fMRI Data Analysis Pipeline Overview.** The pipeline consists of four major stages: (1) Experimental Design (2.1), including participant recruitment and stimulus presentation protocols; (2) Data Acquisition and Preprocessing (2.2–2.3), encompassing slice timing correction, motion correction, co-registration, LSS-based response extraction, and ROI transformation procedures; (3) Confound Control and Cross-Validation Optimization (2.6–2.7), featuring novel ILP methods to balance trial distributions; and (4) Multi-Voxel Pattern Analysis (2.8), combining ROI-based decoding and whole-brain searchlight mapping approaches. While stages 1-2 were primarily implemented by [23], the ROI transformations and stages 3-4 represent novel contributions of the current work.

2.1 Participants and Experimental Design

A total of 47 participants were recruited to investigate the neural correlates of associative learning between CS and US. Of these participants, 43 completed 128 trials across two separate sessions, and the remaining 4 participants completed 64 trials in a single session. This design allowed for investigation of how participants adapt to frequent changes in CS-US contingencies.

The CS consisted of fractal images, and the US consisted of two intensities of painful heat and two intensities of loud sound stimuli. The intensities were adjusted to be clearly perceptible but tolerable to minimize discomfort. The study design follows the protocol described in Horing and Büchel (2022) [23], which focuses on rapid associative learning and frequent unannounced reversals of CS-US contingencies.

2.1.1 Calibration Phase

All participants underwent a 15-minute calibration phase prior to the main experimental sessions (see Figure 2.2A). During this phase, the intensities of both painful heat and loud sound stimuli were calibrated to each participant's individual sensory thresholds, aiming to achieve stimuli that were perceptible yet tolerable. The heat stimuli were delivered via a thermode placed on the participant's forearm, and the sound stimuli were presented through headphones (see Figure 2.2B). This calibration step ensured consistency in stimulus intensity across all trials while minimizing participant discomfort.

2.1.2 Experimental Sessions

Each session comprised 64 trials and included eight unannounced reversals of CS-US contingencies (Figure 2.2A). These reversals were introduced to challenge participants' abilities to relearn associations and thus maintain engagement throughout the experiment. During each session, fMRI and electrodermal activity (EDA) were recorded to capture neural and psychophysiological responses associated with updating conditioned expectations.

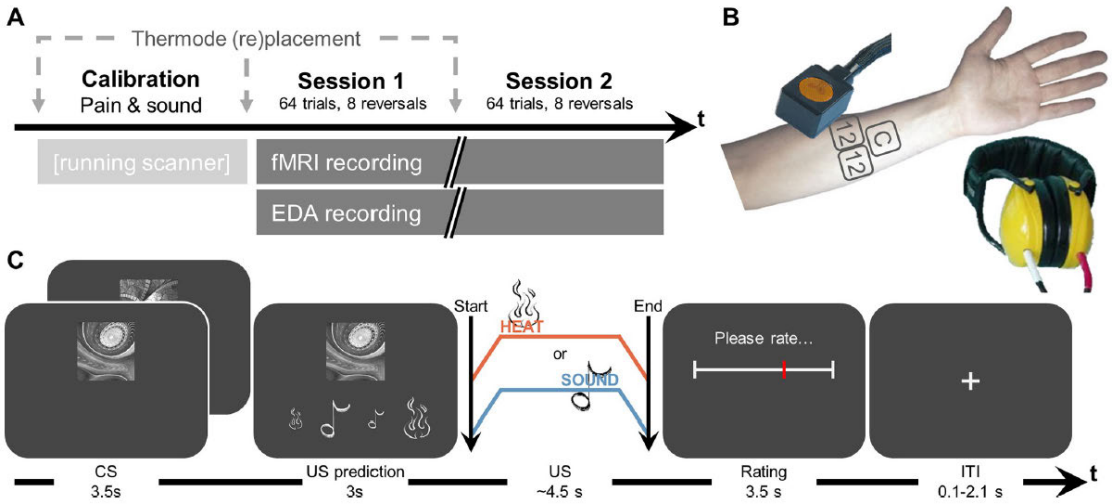

Figure 2.2: **Overview of the experimental protocol.** (A) Calibration phase, session structure, and fMRI/EDA recordings. (B) Thermode placement for heat stimuli and headphone setup for sound stimuli. (C) Trial sequence, showing progression from CS presentation to US prediction, US experience, rating, and intertrial interval. Adapted from [23].

2.1.3 Trial Structure

Each trial followed a consistent sequence (see Figure 2.2C):

1. **CS Presentation (3.5 seconds):** A fractal image, serving as the CS, was displayed.
2. **US Prediction (3 seconds):** Participants were prompted to predict the upcoming US (heat or sound), enabling examination of conditioned expectation formation.
3. **US Presentation (∼4.5 seconds):** The actual US was delivered (painful heat or loud sound), allowing direct comparison between expected and actual outcomes.
4. **Rating (3.5 seconds):** Participants provided a subjective rating of their sensory experience.
5. **Intertrial Interval (0.1–2.1 seconds):** A brief interval separated consecutive trials, permitting stabilization of neural signals before the subsequent CS presentation.

2.1.4 Stimulus Design and Reversal Paradigm

The design employed a deterministic CS-US association paradigm with frequent, unannounced reversals to investigate associative learning (Figure 2.3). The CS comprised a selection of fractal images (Figure 2.3A), two of which were chosen per participant to ensure distinct, individualized representations. The US could be either high or low intensity for both painful heat and loud sound modalities, thereby covering four possible outcomes.

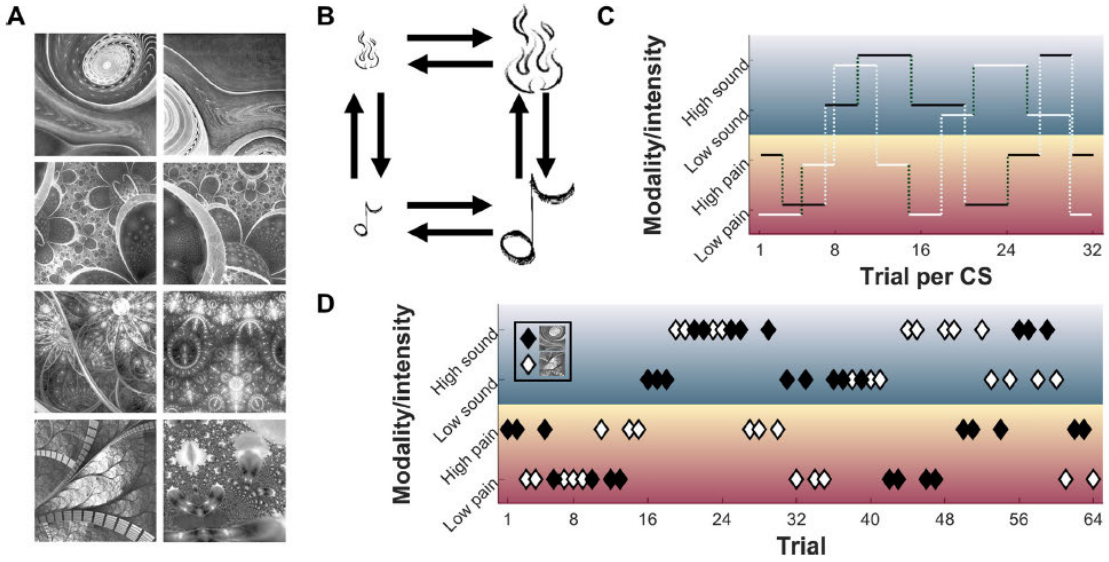

Figure 2.3: **Design of the learning protocol.** (A) Fractal images used as CS. (B) Modality and intensity transitions showing allowable changes and prohibited switches. (C) Structured sequence of trials per CS, indicating the order of presentations. (D) Distribution of trials across the entire session with detailed pairing of US modalities and intensities. Adapted from [23].

Each CS was presented for 32 trials per session, alternating between high or low sound and high or low pain (Figure 2.3C). The reversals occurred on average after 3.75 CS-US associations, thereby requiring participants to adapt continuously and relearn the associations.

This setup was designed to induce rapid associative learning and continuous re-learning under frequent contingency reversals. While such a design challenges the sustained formation of stable conditioned expectations, it specifically facilitates the study of how the brain encodes ongoing changes in predicted outcomes and processes corresponding pre-

diction errors. By periodically altering whether a particular CS indicates painful heat or loud sound, the paradigm compels participants to update their expectations repeatedly. This frequent switching permits a detailed examination of how the brain generates, updates, and transiently encodes aversive expectations—even if these representations do not fully stabilize. Consequently, the paradigm potentially allows for disentangling neural mechanisms of both short-lived expectation states and moment-by-moment prediction error signaling.

2.2 Data Acquisition

This work uses data originally collected and described by Horing and Büchel (2022) [23]. High-resolution functional and anatomical MRI data were acquired using a PRISMA 3T MRI scanner (Siemens, Erlangen, Germany) equipped with a 20-channel head coil. The scanning parameters are restated below.

Functional scans were acquired using a $T2^*$-weighted gradient echo-planar imaging (EPI) sequence with the following parameters:

- **Number of Slices**: 56 transversal slices
- **Slice Thickness**: 1.5 mm
- **Repetition Time (TR)**: 2001 ms
- **Echo Time (TE)**: 30 ms
- **Voxel Size**: 1.5 mm × 1.5 mm × 1.5 mm
- **Gap Between Slices**: 1 mm
- **Field of View (FOV)**: 225 mm × 225 mm × 84 mm

Structural MRI data were acquired using a T1-weighted magnetization-prepared rapid gradient-echo (MPRAGE) sequence with the following parameters:

- **Voxel Size**: 1 mm × 1 mm × 1 mm
- **Number of Slices**: 240 slices

All MRI protocols followed the guidelines described in the original study.

2.3 Preprocessing

Preprocessing was performed using the Statistical Parametric Mapping (SPM) toolbox, version 12 [15]. While spatial normalization (warping individual brains to a standardized space) was performed for group-level searchlight analyses (see Section 2.8.2), first-level preprocessing was conducted in native space. The following steps were applied to correct for common artifacts and ensure suitable alignment for subsequent analyses:

- **Slice Timing Correction**: Temporal offsets among slices, arising from sequential acquisition within each TR, were corrected through interpolation such that all slices were resampled to a common temporal reference [22, 39, 42].
- **Realignment**: Each volume was initially registered to the first volume, and then all volumes were realigned to the mean of those realigned volumes [2, 16, 39]. This minimized effects of head motion by creating a consistent spatial reference.
- **Co-registration**: Each participant's mean functional image (averaged across their realigned volumes) was co-registered to their high-resolution T1 anatomical image. Spatial normalization parameters are estimated with higher precision on the T1 image due to its superior contrast and resolution, and these parameters are subsequently applied to the functional data. The precise alignment between functional and structural images enables accurate normalization [2, 39].
- **ROI Transformation and Thresholding**: The insula ROIs were generated from probabilistic atlases [13, 18, 48] in Montreal Neurological Institute (MNI) space. Inverse deformation fields[1] were applied to transform these probabilistic masks to each participant's native space. Voxels with atlas-based probability values ≥ 15 (indicating presence in at least 50% of the atlas subjects) were retained, ensuring alignment consistency and anatomical specificity (see Appendix A.1.4 for details).

Further details on preprocessing steps can be found in Appendix A.1.

[1] Inverse deformation fields are mathematical transformations that map standardized brain space (like MNI space) back to an individual subject's native brain space. While forward deformation fields warp individual brains to match a standard template, inverse deformation fields do the opposite - they describe how to "unwarp" or transform standardized coordinates back to match each person's unique brain anatomy. This is crucial when applying standardized atlas-based ROIs to individual subject data, as it ensures that the ROIs properly align with each subject's actual brain structure.

2.4 Neural Response Extraction

An LSS modeling approach [33] was applied to estimate trial-specific neural responses. LSS creates individual regressors for each trial, thereby yielding single-trial beta maps (maps of β coefficients), which represent the estimated amplitude of neural activity in response to each experimental condition. These beta maps indicate the strength and spatial pattern of brain activation for each trial. This enhances the detection of subtle or transient responses in event-related designs and allows subsequent decoding analyses to use these trial-level estimates. LSS is particularly advantageous when distinct stimuli (e.g., expected vs. actual US) may have overlapping temporal profiles. Further details on LSS procedures appear in Appendix A.2.

2.5 Trial Selection and Data Inclusion Criteria

To increase the validity of the subsequent expected modality analyses, a dedicated trial and subject selection process was implemented:

- **Excluded Trials:**
  - Early trials in each session that lacked prior CS-US conditioning were discarded to ensure that participants had an opportunity to learn and form stable associations.
  - Trials where participants' reported expectations diverged from established CS-US contingencies were omitted, as such discrepancies could signify misunderstanding or lapses in attention.
- **Removed Subjects:**
  - If more than 33% of a participant's trials showed reported expectations diverging from the established CS-US contingencies, the entire dataset for that participant was excluded from further analysis.
- **Included Trials:**
  - All remaining trials in each session where learning and consistent expectations were presumed to be established.

Applying these criteria resulted in the removal of 4 subjects for the expected modality analysis, leaving a sample of 43 participants. Because the actual modality analysis did not rely on reported expectations, these additional exclusion criteria were not necessary, and no subjects were removed for that condition.

2.6 Cross-Validation Design and Optimization

Both the expected modality and the actual modality decoding analyses used 32 trials per participant, a number chosen to optimize the CV design while meeting several key constraints. This number allows for clean division into 4 balanced folds (8 trials per fold), ensures class balance (16 trials each for heat and sound, and 4 trials each per fold), and represents a practical compromise that allowed most participants to contribute sufficient valid trials after applying exclusion criteria while enabling optimization constraints such as counterbalancing of potential confounding variables.

2.6.1 Initial Design and ILP Optimization

Initial CV was implemented by randomly splitting the 32 trials into four balanced folds, ensuring an equal number of heat and sound trials per fold. However, because the experimental paradigm involved frequent contingency reversals, there was concern that certain trial-level variables (e.g., trial number, session number, CS index) might systematically align with the main labels, potentially producing confounding effects [19, 46].

To address these potential confounds, an ILP approach was employed. In this scheme, each trial was assigned to one of the four folds while imposing additional constraints that balanced potentially confounding variables (e.g., session number, trial number) across the folds [11, 34]. By doing so, it was intended that the class labels (heat vs. sound) would not be inadvertently correlated with the temporal or session-related structure of the data. For a detailed mathematical formulation of the ILP problem, please see Appendix A.3.

2.6.2 Implementation and Outcome

The PuLP optimization library in Python [31] was used to solve the ILP formulation. This yielded folds in which heat and sound trials were distributed evenly, and key variables such as session number and trial order were balanced within each fold. However,

not all participants were able to meet these constraints simultaneously, as some did not have enough valid trials to fulfill the 8-trial-per-fold design once other inclusion criteria had been applied.

As a result of the mismatch-based exclusion criteria (Section 2.5), the expected modality analysis began with 43 participants (since 4 were removed for having more than 33% mismatched trials). Of these 43, 6 could not fulfill the 8-trial-per-fold requirements under the ILP-optimized CV design, leaving a final total of 37 for expected modality. In contrast, the actual modality analysis retained all 47 initial participants (no mismatch-based exclusions apply), but 8 failed to meet the ILP constraints, resulting in a final sample of 39. In both cases, each retained participant had a valid 4-fold assignment of 32 total trials, satisfying class-balance requirements and mitigating temporal/session-related confounds.

2.6.3 Assessing Trial Distribution with Manhattan Distance

After generating the CV folds, an objective measure was required to evaluate how well these designs balanced trial order. The Manhattan distance was therefore employed as an indicator of how interleaved the two labels (e.g., heat vs. sound) were over time. In cases where one label's trial indices are denoted by $\mathbf{P} = [P_1, \ldots, P_m]$ and the other label's trial indices by $\mathbf{Q} = [Q_1, \ldots, Q_m]$, the Manhattan distance D is computed as:

$$D = \sum_{i=1}^{m} \big| P_i - Q_i \big|. \tag{2.1}$$

Because the full derivation and a more detailed example are somewhat lengthy, they are provided in Appendix A.4. Smaller D values indicate that trials of different labels are more evenly interspersed, whereas larger values indicate clustering of identical labels in time.

2.7 Controlling Confounds Using the Same Analysis Approach

The SAA [19] was used to detect and mitigate confounding variables that could invalidate decoding outcomes. This method systematically applies the same pipeline used for the

main research questions to suspected confounders, such as trial order, session identity, CS index, or previous trial modality.

Naive CV designs (Section 2.6) initially served as a baseline. Each potential confounder was treated as a label in classification analyses, using the same preprocessing steps, model setup, and performance metrics. Any significant decoding of these confounders indicated that they might bias results if not controlled. For confounders found to be decodable, analyses were conducted to examine their relationship with the primary labels (expected or actual modality). For instance, the predictability of heat vs. sound purely from trial number was inspected. If trial number predicted the label in any subset of folds, it could drive spurious decoding outcomes, highlighting the need for CV balancing.

To break potential correlations between confounders and the primary labels, an ILP-based CV optimization (Section 2.6.1) was implemented. Following this optimization, suspected confounders were re-tested as classification targets to evaluate the impact of the new design. Meanwhile, the primary labels were again examined to ensure that any observed decoding performance reflected genuine signals rather than residual artifacts.

2.8 Multi-Voxel Pattern Analysis

MVPA was conducted to determine whether distributed neural activity could discriminate heat vs. sound trials under two experimental conditions: (1) the expected modality and (2) the actual modality. In both cases, classification proceeded on single-trial betamaps, and CV folds were optimized to minimize confounding effects as described in Sections 2.4–2.7.

2.8.1 ROI-Based Decoding

For ROI analyses, probabilistic maps of the insular cortex [13, 18, 48] were used (see Section 2.3 for details on ROI transformation). Single-trial betamaps, derived via an LSS approach, were submitted to a Support Vector Machine (SVM) classifier with a radial basis function (RBF) kernel, implemented using The Decoding Toolbox (TDT) [21]. A four-fold CV scheme balanced heat and sound trials while also mitigating trial-order and session-related confounds (Sections 2.6–2.7). Within each insular subregion,

classifier performance was quantified as accuracy minus chance level (i.e., percentage points above or below 50%).

Group-level assessments of these ROI-based decoding accuracies were carried out using prevalence inference. In accordance with the recommendations by Allefeld et al. (2016) [1], 10,000 permutations were performed at the first level for each subject. Subsequently, 1,000,000 second-level permutations were conducted to establish a robust null distribution for the prevalence inference. The significance level (α) was set to 0.05, and the threshold prevalence (γ_0) was set to 0.5, indicating that at least half of the sample needed to demonstrate above-chance decoding for the prevalence null to be rejected. This procedure estimates how many participants exceed chance classification, providing insight into whether insular representations for each modality are widespread across the sample or confined to a smaller subset of individuals. For a more thorough explanation of prevalence inference, see Appendix A.5.

2.8.2 Whole-Brain Searchlight Analysis

A whole-brain searchlight analysis was also performed to identify any additional regions that might encode modality-specific information. As with the ROI analyses, the same set of single-trial betamaps and ILP-optimized CV folds was used. In each spherical neighborhood (radius = 10 mm), the SVM classified heat vs. sound trials, producing an accuracy map for each participant. These maps were then spatially normalized and entered into voxel-wise one-sample t-tests against chance (50%), with significance determined via family-wise error (FWE) correction. For a more detailed explanation of the searchlight procedure, see Appendix A.6.

3 Results

This chapter presents the key findings from applying MVPA to decode expected and actual aversive stimuli from fMRI data. The results are organized into four main sections: First, a systematic investigation of potential confounding variables and their control through optimized CV designs is presented. Second, a detailed evaluation compares the effectiveness of naive versus optimized CV approaches in balancing trial distributions. Third, prevalence inference analyses examine whether decoded neural representations generalize across the population. Finally, while the previous analyses focus on predefined ROIs, particularly the insula, a whole-brain searchlight mapping approach is employed to ensure no brain regions carrying relevant information are overlooked. Throughout these analyses, particular attention is paid to distinguishing genuine neural patterns from potential methodological artifacts.

3.1 Confounder Identification and Control

To accurately decode the expected and actual modalities from neural data, it was crucial to ensure that the results were not influenced by confounding variables that could artificially affect decoding performance. Potential confounders such as the temporal ordering of trials (trial number), session identity (session number), conditioned stimulus identity (CS Index), and previous trial modality can introduce systematic biases, leading to spurious decoding results. Consequently, a systematic approach was undertaken to both identify these potential confounders and implement strategies to control for them, assessing decoding performance differences before and after controlling for these variables.

Initially, decoding analyses were performed using naive CV designs, which were pseudo-randomly created to ensure class balance while partitioning the data into folds without explicitly counterbalancing for potential confounders. These naive designs provided a

baseline assessment of decoding performance and revealed whether the suspected confounders could themselves be decoded from the neural data. To reduce variability introduced by session effects and to simplify the initial analyses, the naive CV designs for early vs. late trials, previous trial modality, and CS Index decoding were restricted to trials from the first session. This restriction ensured that decoding results reflected patterns specific to these variables without interference from broader session-related confounds.

Upon identifying significant decoding of these confounders—indicating their potential to bias the results—CV designs were optimized to control for them. Specifically, trial number and session number were counterbalanced within each fold of the CV, effectively breaking any associations between these confounders and the labels to be decoded. By comparing decoding accuracies before and after controlling for confounders, it was possible to assess the true neural decoding capabilities for the expected and actual modalities and determine whether observed decoding performances were attributable to meaningful neural patterns or merely artifacts resulting from confounding variables.

Through this systematic process of confounder identification and control, the validity of the decoding analyses was enhanced. By ensuring that the findings accurately reflect underlying neural processes associated with the expected and actual sensory modalities, the conclusions drawn from this study were strengthened.

3.1.1 Initial Decoding Performance Across Regions

An inverse brain mask, composed of non-brain voxels, was used as a control to ensure that significant decoding results were not due to systematic artifacts or noise. By applying the decoding analysis to this non-brain region, it was confirmed that observed performance in actual brain regions reflected genuine neural patterns rather than spurious signals.

Figure 3.1 illustrates the decoding accuracy-minus-chance percentages for each variable and label across the different brain regions. The analysis employed a 4-fold naive CV design with balanced trials per class but without additional controls for potential confounds. Decoding performance was evaluated using an SVM with an RBF kernel, and statistical significance was assessed using both uncorrected and Bonferroni-corrected p-values. Mean decoding accuracies are represented by blue dots.

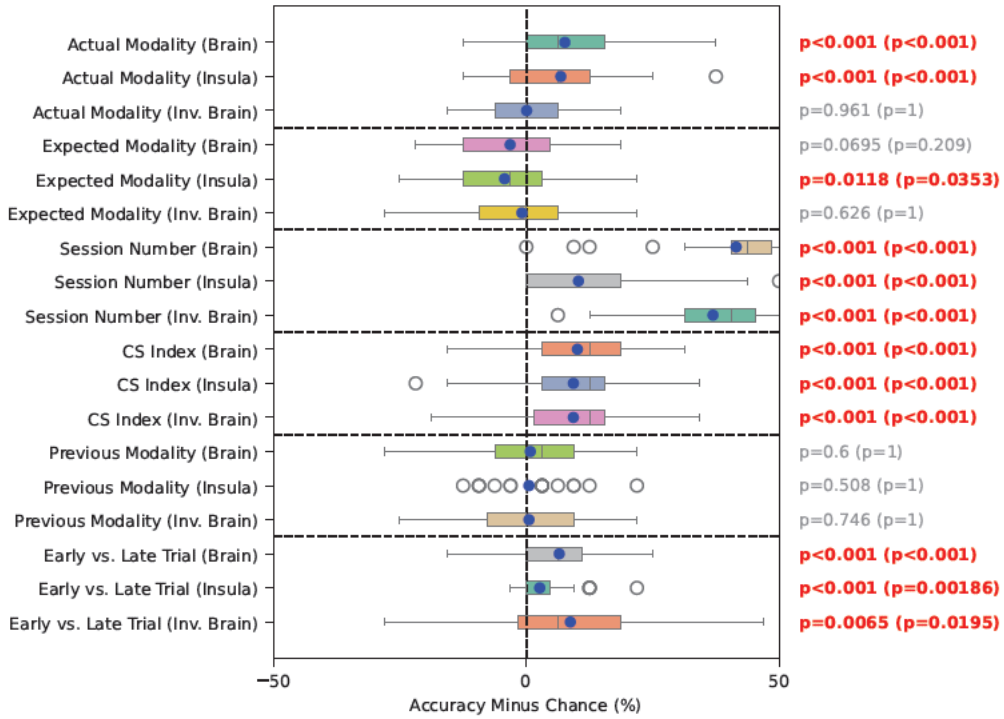

Figure 3.1: **Initial Decoding Results for Confounders and Labels.** This boxplot illustrates the distribution of decoding accuracy-minus-chance percentages for suspected confounding variables (e.g., CS Index, session number, trial number, and previous modality) and primary labels (actual and expected modalities). Results are shown for the whole brain, insula, and an inverse brain mask as a control region. Black p-values indicate uncorrected significance ($p < 0.05$), while red p-values denote Bonferroni-corrected significance. Blue dots represent the mean decoding accuracy for each variable. The analysis is based on a naive CV design aimed at identifying potential confounds and establishing baseline decoding performance.

3.1.2 Summary of Initial Decoding Results

Findings from the naive decoding analyses are summarized as follows:

- **Actual Modality (Label)**: Decoding of the actual modality demonstrated significant results in the whole brain and insula (Bonferroni-corrected $p < 0.001$). The inverse brain mask did not show significant decoding, indicating that the observed performance reflects meaningful neural patterns.

- **Expected Modality (Label)**: Decoding results for the expected modality did not reach above-chance significance in any region. However, in the insula, decoding accuracy was significantly below chance (Bonferroni-corrected $p < 0.05$). This suggests that systematic biases or confounding effects may have influenced decoding performance, leading to below-chance results in this region.
- **Trial Number (Early vs. Late Trials)**: Decoding of early vs. late trials was significant across all regions. This outcome is likely due to scanner-related effects, such as drift and instability over the course of a session, and subject-related factors like fatigue. These results indicate that trial number should be counterbalanced in future analyses to control for its confounding influence.
- **Session Number**: Decoding of session number was significant in the whole brain, insula, and inverse brain mask (Bonferroni-corrected $p < 0.001$). The significant results in the inverse brain mask are consistent with expectations, as scanner noise can vary between sessions. These findings highlight that session effects may act as potential confounds, necessitating counterbalancing of session number in subsequent analyses.
- **Conditioned Stimulus Identity (CS Index)**: Significant decoding was observed for CS Index in the whole brain and insula (Bonferroni-corrected $p < 0.001$). However, similar decoding results in the inverse brain mask suggest that scanner or analysis artifacts may contribute to these findings. This indicates that CS Index could act as a confounding variable and should be investigated further.
- **Previous Trial Modality (Previous Modality)**: Decoding results for previous modality were not significant across regions. However, given the potential influence of trial number, it was considered that trial number might also act as a confounder here and warranted further investigation.

Based on these results, trial number, session number, and CS Index emerged as potential confounders affecting decoding performance. Given trial number's significant effects across all regions, the next step was to investigate whether trial number alone could predict other variables—specifically CS Index (which showed similar artifacts) and previous modality (which, despite non-significant decoding, might still be influenced by trial ordering). This would help isolate trial number's specific role in confounding the results.

3.1.3 Investigating Trial Number as a Confounding Variable

To determine the extent to which trial number acts as a confounder, analyses were conducted where trial number was used as the sole input feature for decoding CS Index and previous modality. By isolating trial number, it was possible to assess whether time-related factors within a session—such as scanner drift or participant fatigue—are correlated with these variables and could therefore be decoded by trial number alone.

It is important to note that a simple correlation coefficient between trial number and the other variables would not suffice for this analysis. Correlations could vary across folds of the CV design, potentially canceling each other out when averaged. Despite this, such fold-specific correlations could still influence decoding performance, making it essential to assess the relationship through the decoding framework rather than relying on correlation metrics alone.

Figure 3.2 presents the decoding accuracy-minus-chance percentages for CS Index and previous modality when trial number is the only input feature. Results are compared between naive CV designs and trial number-optimized designs, where data are counterbalanced for trial number.

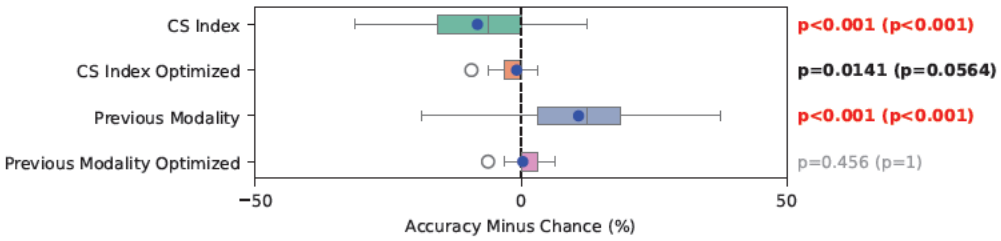

Figure 3.2: **Decoding Results of CS Index and Previous Modality Using Trial Number as the Only Input Feature.** Comparison of decoding results between naive and optimized CV designs, where trial number was the sole input feature for the SVM. The boxplot shows decoding accuracy-minus-chance percentages for CS Index and previous modality. Uncorrected (black) and Bonferroni-corrected (red) p-values indicate statistical significance. Blue dots represent mean decoding accuracy.

In the naive designs, decoding performance for both CS Index and previous modality significantly deviated from chance level (both uncorrected and Bonferroni-corrected p-values were significant). This indicates that trial number is associated with these variables within the CV design, leading to significant decoding. Such associations suggest that the

observed decoding results may be influenced by trial-related confounding effects rather than reflecting genuine neural representations.

When trial number-optimized designs were used, where trial numbers were counterbalanced, decoding performance dropped to chance levels and was no longer significant—except for uncorrected significance in CS Index decoding, likely due to imperfect counterbalancing for certain subjects. This demonstrates that counterbalancing trial numbers effectively broke the associations, potentially allowing for a more accurate assessment of the neural representations of CS Index and previous modality.

3.1.4 Decoding CS Index and Previous Modality with Optimized CV Designs

Given the identified confounding effects of trial number, decoding analyses for CS Index and previous modality were repeated using optimized CV designs that counterbalanced trial number within each fold.

Figure 3.3 shows the decoding results using these optimized designs across the whole brain, insula, and inverse brain mask.

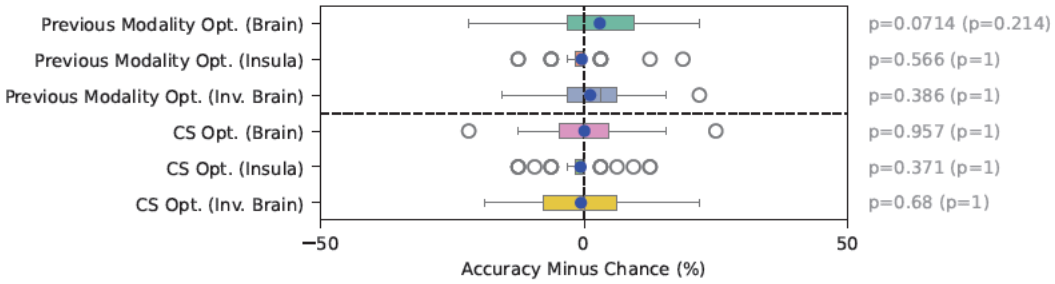

Figure 3.3: **Decoding Results of CS Index and Previous Modality for Optimized CV Designs.** This boxplot illustrates decoding accuracy-minus-chance percentages for CS Index and previous modality using optimized CV designs. Blue dots represent mean decoding accuracy. Uncorrected (black) and Bonferroni-corrected (red) p-values indicate statistical significance.

The results indicated that decoding performance for both CS Index and previous modality was no longer significant in any region. This suggests that the significant decoding observed in the naive analyses was largely due to trial number effects rather than genuine neural patterns associated with these variables.

3.1.5 Decoding Expected and Actual Modality with Optimized Designs

Based on the results of the confounder identification, decoding analyses were conducted using CV designs optimized to counterbalance both session number and trial number. This approach aimed to determine whether the expected and actual modalities could be decoded independently of these confounding variables and to evaluate whether decoding performance improves as a result of the optimization.

Figure 3.4 illustrates the decoding accuracy-minus-chance percentages for expected and actual modalities using the optimized designs.

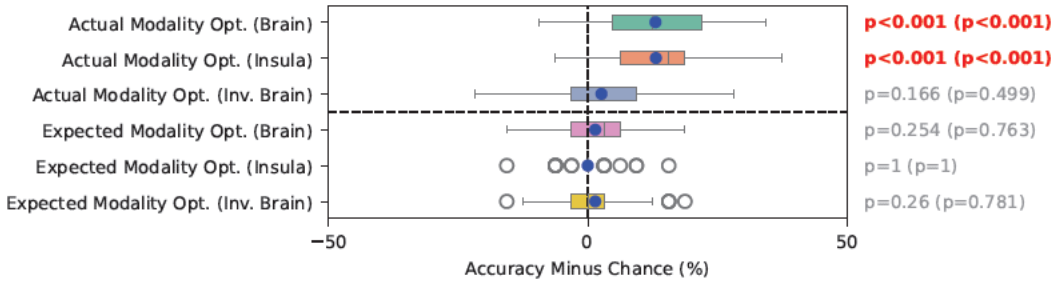

Figure 3.4: **Decoding Results of Expected and Actual Modality for Optimized Designs.** This boxplot shows decoding accuracy-minus-chance percentages using CV designs optimized by counterbalancing session number and trial number. Blue dots represent mean decoding accuracy. Uncorrected (black) and Bonferroni-corrected (red) p-values indicate statistical significance.

For the expected modality, decoding performance did not reach statistical significance in any region. However, mean decoding accuracy increased compared to the naive design and was no longer negative, suggesting some improvement in decoding after controlling for session and trial number. Despite this, the expected modality could not be reliably decoded from neural data.

In contrast, decoding of the actual modality not only remained significant in both the whole brain and insula regions (Bonferroni-corrected $p < 0.001$) but also showed a substantial improvement in decoding accuracy compared to the naive design. This demonstrates that neural representations of the actual sensory modalities are robust and distinct enough to be reliably decoded. Moreover, it highlights that controlling for confounders not only maintained but also improved decoding performance.

3.1.6 Final Validation Using Session and Trial Number as Input Features

To further validate the confound control measures, an analysis was conducted using session number and trial number individually as the sole input features for decoding the expected and actual modalities. This analysis aimed to confirm that the counterbalancing optimization effectively broke the association between the confounding variables and the labels within the CV folds.

Figure 3.5 presents the decoding results comparing naive and optimized CV designs.

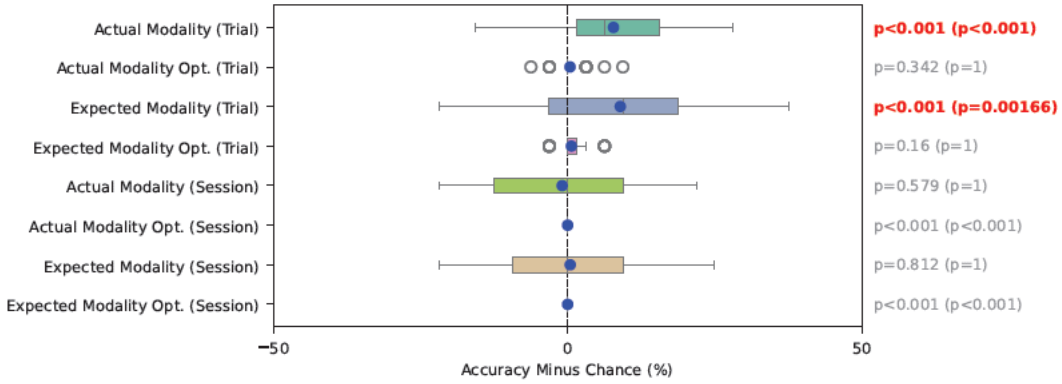

Figure 3.5: **Decoding Results Using Session or Trial Number as Input Features.** This boxplot illustrates decoding performance when session number or trial number is used individually as the sole input feature. Blue dots represent mean decoding accuracy. Uncorrected (black) and Bonferroni-corrected (red) p-values indicate statistical significance.

When session number was used as the input feature, decoding performance did not reach significance in the naive design. However, the high variance across subjects indicated that, in many cases, there was an association between the respective label and session number. This association led to either above- or below-chance decoding performance, thereby influencing the decoding of the label. In the optimized design, variance was reduced, and decoding performance was uniformly at chance levels, showing that the optimization effectively broke the association.

When trial number was the input feature, decoding performance was significantly above chance in the naive design for both expected and actual modalities, confirming its role as a confounder. In the optimized design, decoding accuracy returned to chance levels, demonstrating that the optimization effectively eliminated trial number as a confounding

factor. However, some minor variance remained, as the experimental design did not allow for perfect counterbalancing of trial numbers.

3.2 Evaluation of CV Designs: Optimized vs. Naive Approaches

This section examines how effectively the CV optimization process balanced trial distributions across folds, taking into account trial number and assigned labels (e.g., Heat and Sound). Although the confounder analysis revealed that the optimization substantially reduced the predictive influence of trial number, some residual correlation between trial numbers and their respective labels persisted in the CV designs (see Figure 3.5). The section provides a visual and quantitative demonstration of the improvements achieved by comparing representative examples of individual subjects and overall group metrics.

3.2.1 Actual Modality Designs

Figure 3.6 presents two representative subjects under naive (top plot) and optimized (bottom plot) CV designs. The upper plot (subject 12) exhibits a high Manhattan distance, indicating noticeable trial clustering in the naive design. By contrast, the lower plot (subject 1) demonstrates substantially more uniform trial balancing with a minimal Manhattan distance. Different folds are depicted using distinct colors, and a vertical line after trial 64 separates the two sessions.

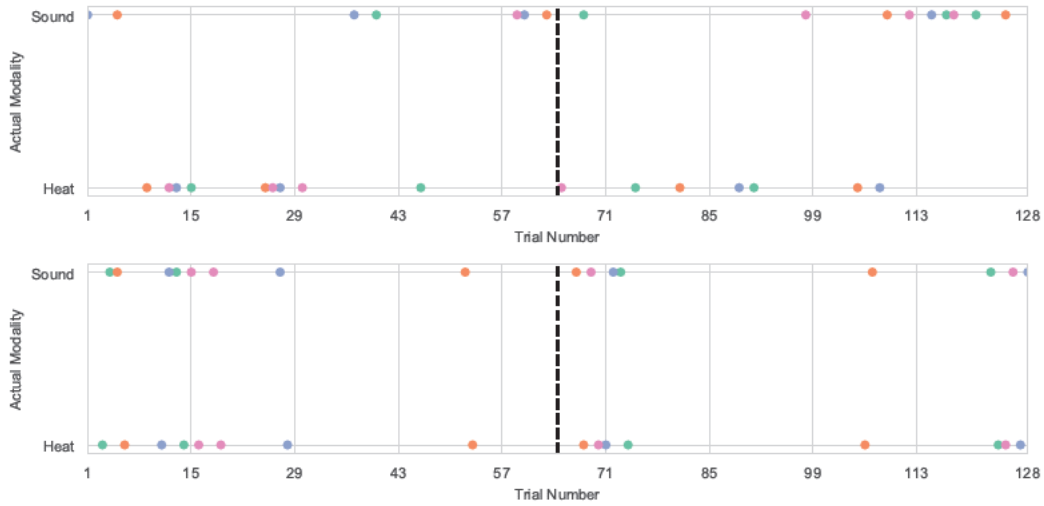

Figure 3.6: **Visual representation of label distributions across sessions and trial numbers for selected subjects for actual modality decoding.** The upper plot shows the subject with the highest Manhattan distance from the naive CV design (subject 12), while the lower plot shows the subject with the lowest Manhattan distance in the optimized CV design (subject 1, with a minimum distance of 16 shared by multiple subjects). Data points are color-coded by CV folds, and the vertical line between trial numbers 64 and 65 marks the start of session 2.

3.2.2 Expected Modality Designs

Figure 3.7 highlights two subjects for the expected modality. The top plot (subject 28) reveals considerable imbalances under the naive approach, whereas the bottom plot (subject 3) achieves more uniform labeling after the optimized procedure. This outcome underscores the consistent effectiveness of the optimization in reducing temporal and labeling confounds across folds.

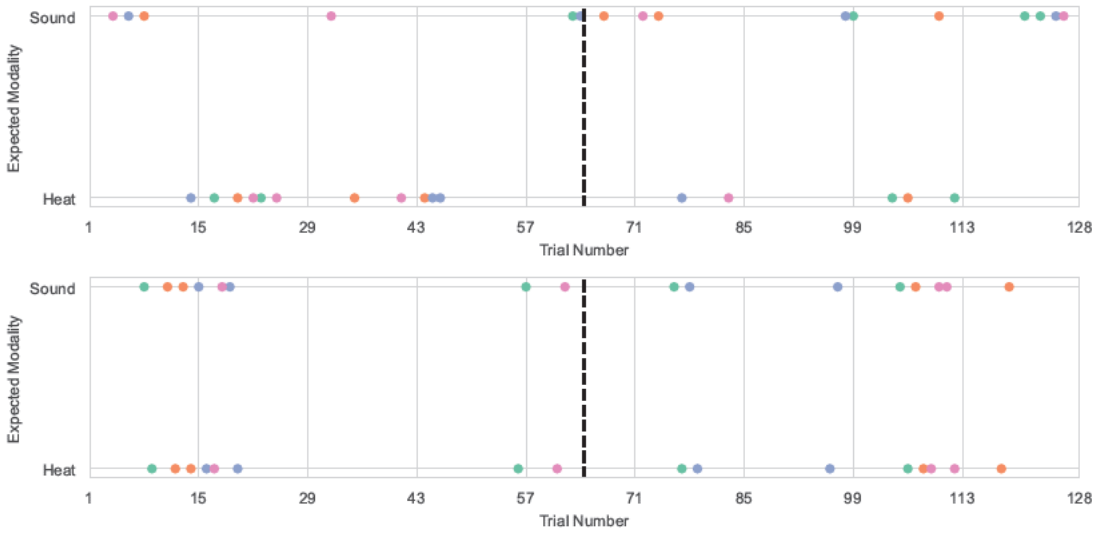

Figure 3.7: **Visual representation of label distributions across sessions and trial numbers for selected subjects for expected modality decoding.** The upper plot shows the subject with the largest Manhattan distance from the naive CV design (subject 28), while the lower plot represents the subject with the lowest Manhattan distance in the optimized CV design (subject 3, with a minimum distance of 16 shared by multiple subjects). Data points are color-coded by CV folds, and the vertical line between trial numbers 64 and 65 marks the start of session 2.

3.2.3 Comparison of Naive vs. Optimized Across All Subjects

Figure 3.8 provides an aggregated overview of Manhattan distances for all subjects and both modalities (actual and expected). The naive designs produce larger distance values, indicating a higher degree of clustering. By contrast, the optimized designs consistently lower these distances, reflecting a more balanced trial distribution.

Notably, the distances for the optimized CV designs in the expected modality are often higher than for the actual modality. As discussed in Section 2.5, this can be attributed to the removal of certain trials in the expected modality condition, which leads to fewer data points per subject and potentially less stable trial balancing. The reduced sample size may also explain the relatively higher variance observed, for example, in Figure 3.5.

Overall, the substantial gains observed across subjects and modalities confirm the efficacy of the optimization procedure in mitigating trial-number imbalances.

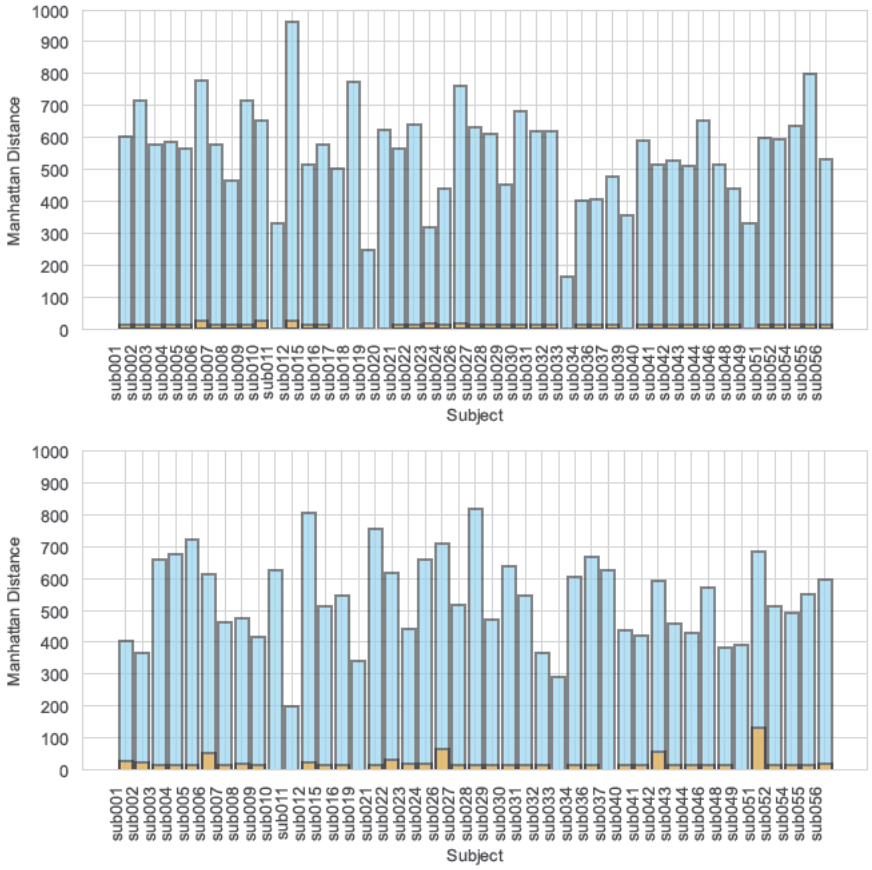

Figure 3.8: **Comparison of Manhattan distances for trial number distributions across naive and optimized CV designs for actual and expected modality decoding.** The top plot represents the actual modality, while the bottom plot illustrates the expected modality. Each bar corresponds to the Manhattan distance for individual subjects, with overlays highlighting differences between naive (blue) and optimized (orange) conditions. The results demonstrate the superior balance achieved with optimized designs across the dataset.

3.3 Insula Population-Level Analysis Through Prevalence Inference

To evaluate the degree to which decoding of actual and expected aversive stimuli is representative at the population level, a prevalence inference analysis was conducted on insular regions [13, 18, 48]. This method not only tests whether at least one participant shows above-chance decoding ("global null") but also indicates the proportion of partic-

ipants in the population for which this effect holds. Tables 3.1 and 3.2 summarize the key parameters across insular subregions and for the entire insula. The meaning of each column is as follows:

- **puGN / pcGN**: Uncorrected (u) and corrected (c) p-values under the **global null hypothesis (GN)**. Rejection of the global null indicates that some subjects in the population have an above-chance effect.
- **puMN / pcMN**: Uncorrected (u) and corrected (c) p-values using the **minimum statistic (MN)** approach, testing against the global null but with a more conservative procedure.
- **gamma0c** (γ_{0c}): The **prevalence threshold** or lower bound on the fraction of the population that exhibits above-chance decoding. A higher value implies that a substantial proportion of individuals show the effect.

3.3.1 Actual Modality

Most individual insular subdivisions did not yield robust prevalence estimates beyond rejecting the global null (Table 3.1), suggesting that above-chance decoding may be restricted to a smaller subset of participants for those specific subregions. However, the analysis of the entire insula revealed notably small p-values under both GN and MN formulations (e.g., $p_{\text{uGN}} = 0.00002$, $p_{\text{cGN}} = 0.00019$). Additionally, the prevalence threshold γ_{0c} reached approximately 0.69446, indicating that fewer than about 31% of individuals can be confidently assumed not to show above-chance decoding. Collectively, these findings imply a population-typical neural encoding of the actual modality within the insula.

Table 3.1: Prevalence Results for Actual Modality Across ROIs

ROI	puGN	pcGN	puMN	pcMN	gamma0c
Posterior Long Gyrus (L)	0.01386	0.16410	0.124840	0.268453	–
Posterior Long Gyrus (R)	0.01571	0.16410	0.132462	0.274825	–
Anterior Short Gyrus (L)	0.56733	0.99968	0.753989	0.999921	–
Anterior Short Gyrus (R)	0.55432	0.99968	0.745358	0.999919	–
Middle Short Gyrus (L)	0.97399	1.00000	0.986912	1.000000	–
Middle Short Gyrus (R)	0.31324	0.97405	0.562101	0.988637	–
Posterior Short Gyrus (L)	0.54022	0.99968	0.735890	0.999915	–
Posterior Short Gyrus (R)	0.09844	0.66722	0.319202	0.773444	–
Anterior Inferior Cortex (L)	0.75400	1.00000	0.868554	1.000000	–
Anterior Inferior Cortex (R)	0.98996	1.00000	0.994968	1.000000	–
Anterior Long Gyrus (L)	0.53602	0.99968	0.733047	0.999915	–
Anterior Long Gyrus (R)	0.08792	0.66722	0.302184	0.767781	–
Whole Insula	0.00002	0.00019	0.006501	0.006689	0.69446

3.3.2 Expected Modality

Unlike the actual modality, the expected modality showed no conclusive prevalence estimates in any subregion or in the insula as a whole (Table 3.2). Although some subregions displayed relatively low p-values initially—signifying that some participants may decode the expected modality above chance—no region demonstrated a sufficiently consistent effect across subjects to yield a non-trivial γ_{0c}. Thus, there is no strong evidence that above-chance decoding of the expected modality is shared by a substantial proportion of the population.

Table 3.2: Prevalence Results for Expected Modality Across ROIs

ROI	puGN	pcGN	puMN	pcMN	gamma0c
Posterior Long Gyrus (L)	0.37921	0.81778	0.617659	0.930330	–
Posterior Long Gyrus (R)	0.80684	0.99997	0.898375	0.999997	–
Anterior Short Gyrus (L)	0.00942	0.81778	0.104064	0.836742	–
Anterior Short Gyrus (R)	0.95470	1.00000	0.977094	1.000000	–
Middle Short Gyrus (L)	0.82859	1.00000	0.910372	1.000000	–
Middle Short Gyrus (R)	0.39285	0.99997	0.628534	0.999989	–
Posterior Short Gyrus (L)	0.21489	0.99357	0.467088	0.996573	–
Posterior Short Gyrus (R)	0.89142	1.00000	0.944190	1.000000	–
Anterior Inferior Cortex (L)	0.82349	1.00000	0.907573	1.000000	–
Anterior Inferior Cortex (R)	0.46322	0.99357	0.681896	0.997955	–
Anterior Long Gyrus (L)	0.10887	0.81778	0.335196	0.878859	–
Anterior Long Gyrus (R)	0.33462	0.99357	0.580690	0.997304	–
Whole Insula	0.70780	0.99997	0.841631	0.999995	–

3.3.3 Summary of Prevalence Findings

These results, together with the initial decoding results (see Figure 3.4), highlight the pronounced divergence between actual and expected modality representations in the insula. For the actual modality, prevalence inference confirms that not only is there a statistically significant effect at the group level, but it is also widely shared among participants—especially when considering the insula in its entirety. In contrast, the expected modality's signal appears sporadic and insufficiently prevalent to conclude that it reflects a majority-level phenomenon. Consequently, prevalence inference provides a more nuanced population perspective, underscoring that the actual modality decoding effect in the insula is robust and common, whereas the expected modality effect—if present at all—does not generalize broadly across individuals.

3.4 Whole-Brain Searchlight Analysis

To further explore and ensure that no region of potential relevance was overlooked in the spatial distribution of neural representations underlying the decoding of both expected and actual modalities, a voxel-wise searchlight analysis [26] was conducted. This approach aimed to identify localized clusters of voxels throughout the brain carrying information that discriminated between conditions above chance level. Specifically, a spherical cluster of voxels was iteratively centered on each voxel in the brain, and the SVM-based decoding analysis described previously was applied within these local neighborhoods. By systematically evaluating performance across the entire brain volume, searchlight mapping provides a fine-grained characterization of the spatial patterns of neural information encoding.

Each subject's searchlight accuracy maps were generated for both the actual and expected modality comparisons under the optimized CV design, ensuring that trial number and session number were adequately counterbalanced. These individual-level accuracy maps were registered to a common template space and subjected to second-level random-effects analyses. This involved one-sample t-tests against chance-level performance (50%), determining whether any local voxel clusters exceeded chance at the group level.

The resulting t-statistic maps were thresholded using FWE corrections to control for multiple comparisons, ensuring that only robust and consistent patterns of above-chance decoding emerged as significant clusters. In accordance with the prevalence inference

results, the searchlight analysis for the actual modality identified distributed clusters, including regions within the insular cortex, that significantly encoded modality-specific information for a substantial proportion of subjects. These second-level results are illustrated in Figure 3.9, where both FWE-corrected and uncorrected maps are presented. As depicted, two significant clusters emerged for the positive contrast under FWE correction, reinforcing the notion that actual modality information is robustly represented across subjects. For both the positive and negative contrasts, there were a small amount of significant voxels scattered around at the uncorrected level. However, after FWE correction, no significant voxels were observed for the negative contrast.

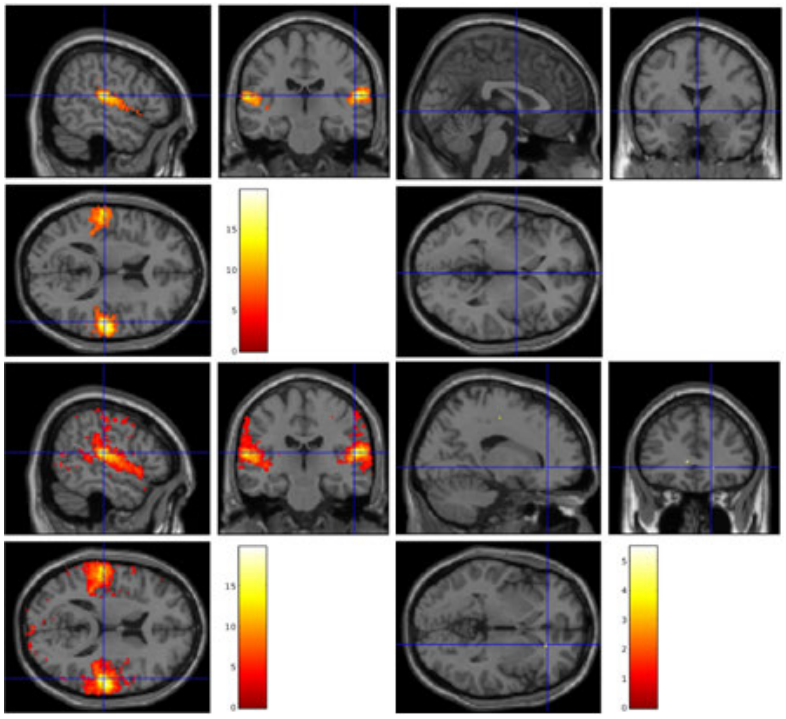

Figure 3.9: **Actual Modality: Second-Level Results.** Heatmaps of t-values for positive (left) and negative (right) contrasts from t-tests. The upper row shows results corrected for FWE, and the lower row shows uncorrected results ($p < 0.001$). The respective crosshair is positioned at the global maximum activation. Brighter colors (red and yellow) represent higher t-values. For the positive contrast, two significant clusters can be observed, while for the negative contrast, only a small amount of significant voxels were evident uncorrected, and none were significant under FWE correction.

In sharp contrast, and consistent with the weak prevalence evidence for the expected modality, no significant clusters emerged at the FWE-corrected level for the expected

modality. Although some scattered voxels surpassed uncorrected thresholds, these did not form reliable clusters at the group level. For both the positive and negative contrasts, there was a small amount of significant voxels scattered around uncorrected. After FWE correction, there were no significant voxels for either contrast. Thus, while some individuals may show neural signatures related to the expected modality, these patterns were not sufficiently robust or consistent across subjects to reach significance after rigorous correction for multiple comparisons. These second-level results are summarized in Figure 3.10, which displays both FWE-corrected and uncorrected maps for the expected modality. As shown, no significant clusters were observed for the positive contrast under FWE correction, while only a small number of significant voxels appeared for the negative contrast at uncorrected thresholds.

The second-level results for the naive CV designs can be found in Appendix A.7, in Figures A.2 and A.3. In these analyses, many highly significant voxels were observed for the negative contrast, while the results for the positive contrast were similar to the findings of the optimized CV designs.

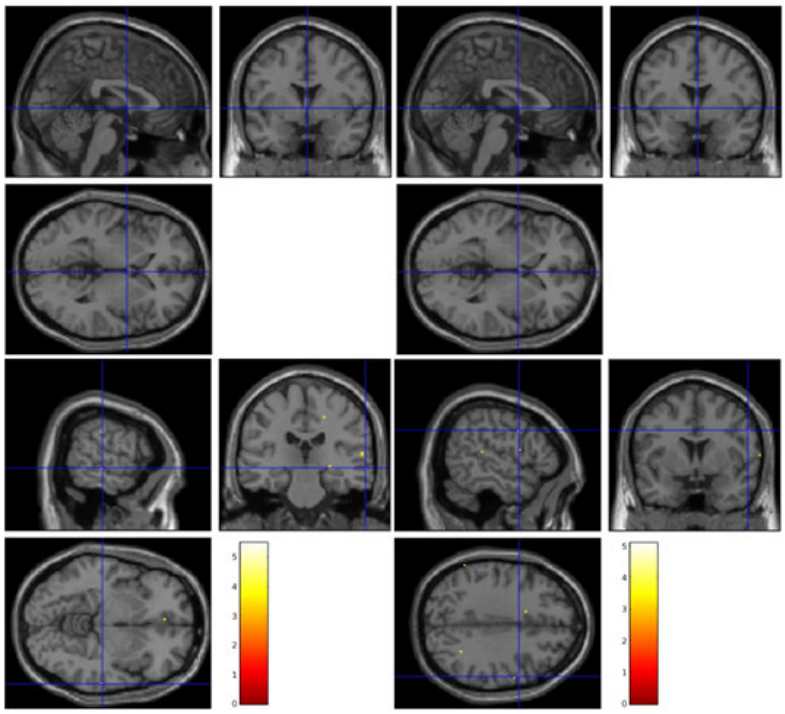

Figure 3.10: **Expected Modality: Second-Level Results.** Heatmaps of t-values for positive (left) and negative (right) contrasts from t-tests. The upper row shows results corrected for FWE, and the lower row shows uncorrected results ($p < 0.001$). The crosshair is positioned at the respective global maximum activation, if present. Brighter colors (red and yellow) represent higher t-values. No significant voxels were observed for the positive contrast under FWE correction, with only a few single voxels appearing when uncorrected. Consistent with this, for the negative contrast, only a small number of significant voxels were observed at the uncorrected level, and none remained significant after FWE correction.

4 Discussion

This thesis investigated the neural representations of expected and actual aversive stimuli during associative learning using fMRI-based MVPA. The results revealed an important difference: while the actual stimulus modality was robustly decodable across subjects neural patterns in the insula did not reliably encode expected stimulus modality—despite clear behavioral evidence that participants learned and maintained accurate predictions throughout the task (compare Section 2.5; only 4 out of 47 participants had more than 33% errors; for a deeper analysis see [23]).

Through methodological advances including the SAA and optimized CV using ILP, this work identified and controlled for temporal confounds that could have led to spurious results. The successful decoding of actual stimuli served as a crucial positive control, validating the analysis pipeline and demonstrating sufficient sensitivity to detect stimulus-related patterns when present. These methodological approaches may help inform the design and analysis of future studies examining neural representations during associative learning.

4.1 Methodological Contributions

This work makes several important methodological contributions to the field of information-based neuroimaging. It was demonstrated how the SAA can systematically reveal and control for temporal confounds that might otherwise remain undetected in MVPA studies. While previous research has highlighted the importance of controlling for confounds in neuroimaging [19, 43, 45, 46], the application of SAA provided a systematic framework for identifying and addressing these issues. By applying the same decoding analyses to potential confounding variables, it was possible to detect significant trial order effects that could have led to spurious below-chance accuracies in the results.

A particularly novel contribution is the development of optimized CV designs using ILP. Traditional approaches to CV in MVPA studies often assume that simple randomization or balanced class distributions are sufficient to prevent bias. However, the results demonstrate that even seemingly balanced designs can harbor systematic confounds related to trial order and session effects. The ILP approach provides a mathematical framework for explicitly controlling these confounds while maintaining the essential properties of CV. This method could be particularly valuable for studies employing complex experimental designs where multiple potential confounds need to be balanced simultaneously.

While the ILP approach proved effective at controlling identified confounds, its implementation came with substantial constraints on the usable data. Although the original dataset contained 128 trials per subject, the requirements for balanced confound control and conditioned expectations led to a dramatic reduction to just 32 trials per subject. This reduction occurred in two stages: first, trials without established conditioned expectations had to be removed to ensure valid expectation-related analyses. Second, the combined requirements of 4-fold CV and balanced trial distributions across confounding variables further limited the number of usable trials. The need to maintain equal numbers of trials across conditions while simultaneously balancing multiple confounding factors meant that only 32 trials per subject could be effectively utilized in the optimized design. Additionally, had the SAA identified more confounding variables requiring control, this would have further constrained the optimization problem, potentially leading to either worse trial number distance metrics (e.g., Manhattan distance) or the exclusion of more subjects who could not meet all balancing requirements simultaneously. This substantial reduction in usable trials highlights an inherent trade-off between rigorous confound control and maintaining adequate trial numbers in MVPA studies.

Prevalence inference methods as proposed by [1] were also employed to assess the population-level generalization of the decoding results. This approach provided a more precise characterization of the findings compared to traditional second-level t-tests on accuracy maps, revealing that actual modality information was consistently represented in a majority of the population, while the absence of decodable expected modality representations reflected a genuine lack of consistent neural patterns across subjects rather than just insufficient statistical power.

4.2 Interpretation of Expected Modality Results

The inability to decode expected stimulus modality from neural patterns warrants careful examination, particularly since behavioral data demonstrated that participants successfully learned and maintained accurate predictions throughout the task. The null result was clearly established through chance-level decoding accuracies and absence of significant prevalence inference results, despite rigorous methodological controls.

Several factors likely contributed to this dissociation between behavioral success and neural decodability. Participants likely learned not only the specific CS-US associations but also to expect frequent reversals, which could add substantial variance to the neural representations of expected stimuli. The limited number of consecutive CS-US pairings (3.75 on average) before each reversal likely impacts the strength of learned associations [40], and may also prevent the development of stable neural patterns that could be detected with current MVPA methods. Additionally, even after controlling for confounds, the remaining variance in the neural signals may have been too high to reliably decode these rapidly changing patterns using current MVPA methods [19, 43].

These results suggest that future work investigating neural representations of expectations should use experimental designs specifically optimized for this purpose, with longer learning periods between reversals to allow more robust expectations to form. While the current design served well for studying prediction error processing [23], different experimental parameters may be needed to capture stable expectation representations.

4.3 Limitations and Future Directions

Several limitations of this study should be acknowledged. First, the rapid reversal design, while effective for studying prediction errors, may have created a trade-off between maintaining participant engagement and allowing sufficient time for decodable expectations to form. Future studies might benefit from designs that incorporate longer learning periods while maintaining participant attention through other means.

The focus on the insula as a ROI represents an additional limitation. While this choice was well-motivated by previous research [12, 23, 24], expectations might be represented in broader neural networks or in dynamic patterns of connectivity that the current analysis

approach could not capture. Future studies might benefit from employing network-based analyses or investigating temporal dynamics of neural patterns during expectation formation.

From a methodological perspective, the ILP approach to CV optimization proved effective at controlling identified confounds, though it required a substantial reduction from the original 128 trials per subject to just 32 trials. However, it cannot be excluded that more complex confounding variables remained undetected by SAA and, therefore, were not accounted for in the optimization. Furthermore, the residual variance observed when using trial number as the sole input feature (see Figure 3.5) suggests that perfect counterbalancing could not be achieved within the constraints of the current experimental design. This is further supported by the variability in Manhattan distances observed in the optimized CV designs (Figure 3.8, orange bars), where some of the achieved distances deviate from the theoretical optimum of 16 for 32 trials. These findings underscore the importance of addressing temporal confounds during the design phase of future studies, rather than attempting to control for them post-hoc through analytical methods alone.

A fundamental limitation of the ILP approach is that it can only practically optimize for a finite number of confounding variables. As the number of confounders increases, the optimization problem becomes increasingly constrained and may eventually reach a point where no perfect solution exists that satisfies all constraints simultaneously. This means that in experimental designs with many potential confounding variables, researchers may need to carefully prioritize which confounders to control for through ILP optimization, potentially leaving some confounding effects unaddressed.

Several promising directions for future research emerge from these findings. Studies specifically designed to investigate expectation formation might employ classical conditioning paradigms with extended learning periods or no reversals at all, potentially revealing expectation-related neural patterns that may have been obscured by the rapid reversals used here. Additionally, multi-session experimental designs would allow for leave-one-session-out CV approaches, providing a more principled way to address temporal confounds. Such designs would naturally control for various dependencies that arise within sessions (scanner drift, fatigue effects, trial order), while allowing participants to develop more robust expectations through repeated exposure to the learning paradigm. This approach could provide a methodologically sound framework for investigating the neural representations of expectations while minimizing the influence of temporal confounds that were identified in the current study.

4.4 Conclusion

This thesis has advanced the methodological understanding of MVPA and provided insights into how the brain processes actual aversive stimuli. Through careful application of MVPA techniques, this work demonstrated robust decoding of actual aversive modalities within the insular cortex, supporting its established role in processing multimodal aversive information. The inability to decode expected modalities, despite successful behavioral learning by participants, suggests complex dynamics in how expectations are neurally represented during associative learning tasks.

The methodological contributions extend beyond the specific experimental context. The development and implementation of the SAA and ILP optimization for CV designs provides the field with robust tools for controlling confounding variables in MVPA studies. These methods offer a systematic framework for improving the reliability of multivariate analyses in neuroimaging research.

The findings have implications for basic research in the field of cognitive neuroscience. They highlight the importance of considering experimental design parameters when investigating neural representations of expectations, and demonstrate how rigorous methodological controls can enhance the interpretation of multivariate pattern analyses. The successful decoding of actual stimulus modalities while failing to decode expected modalities raises important questions about the neural mechanisms of expectation formation and representation.

These results underscore the importance of rigorous methodology in decoding studies while advancing the understanding of how the brain represents and processes aversive experiences. Future work building on these findings and methodological advances will be well-positioned to further unravel the neural mechanisms underlying expectation formation and processing.

Bibliography

[1] Allefeld, C., Görgen, K., and Haynes, J.-D. Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *NeuroImage 141* (2016), 378–392.

[2] Ashburner, J., and Friston, K. Rigid body registration. In *Human Brain Function*, R. Frackowiak, K. Friston, C. Frith, R. Dolan, C. Price, S. Zeki, J. Ashburner, and W. Penny, Eds., 2nd ed. Academic Press, 2003, ch. 2.

[3] Atlas, L. Y. How instructions, learning, and expectations shape pain and neurobiological responses. *Annual Review of Neuroscience 46* (Jul 2023), 167–189. Epub 2023 Mar 14.

[4] Atlas, L. Y., and Wager, T. D. How expectations shape pain. *Neuroscience Letters 520*, 2 (Jun 2012), 140–148. Epub 2012 Mar 21.

[5] Behrens, T. E., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. Learning the value of information in an uncertain world. *Nature Neuroscience 10* (2007), 1214–1221.

[6] Brodersen, K. H., Wiech, K., Lomakina, E. I., Lin, C., Buhmann, J. M., Bingel, U., Ploner, M., Stephan, K. E., and Tracey, I. Decoding the perception of pain from fmri using multivariate pattern analysis. *NeuroImage 63*, 3 (2012), 1162–1170.

[7] Calhoun, V., Adali, T., Pearlson, G., and Pekar, J. A method for making group inferences from functional mri data using independent component analysis. *Human Brain Mapping 14*, 3 (Nov 2001), 140–151.

[8] Clark, L., Cools, R., and Robbins, T. W. The neuropsychology of ventral prefrontal cortex: Decision-making and reversal learning. *Brain and Cognition 55*, 1 (2004), 41–53.

[9] Coll, M.-P., Slimani, H., Woo, C.-W., Wager, T. D., Rainville, P., Vachon-Presseau, É., and Roy, M. The neural signature of the decision value of future pain. *Proceedings of the National Academy of Sciences 119*, 23 (2022), e2119931119.

[10] Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., and Marchal, G. Automated multi-modality image registration based on information theory. In *Proc. Information Processing in Medical Imaging* (Dordrecht, The Netherlands, 1995), Y. Bizais, C. Barillot, and R. Di Paola, Eds., Kluwer Academic Publishers, pp. 263–274.

[11] Conforti, M., Cornuéjols, G., and Zambelli, G. *Integer Programming*. Springer, Cham, Switzerland, 2014.

[12] Craig, A. D. How do you feel–now? the anterior insula and human awareness. *Nature Reviews Neuroscience 10*, 1 (January 2009), 59–70.

[13] Faillenot, I., Heckemann, R. A., Frot, M., and Hammers, A. Macroanatomy and 3d probabilistic atlas of the human insula. *NeuroImage 150* (2017), 88–98.

[14] Fazeli, S., and Büchel, C. Pain-related expectation and prediction error signals in the anterior insula are not related to aversiveness. *Journal of Neuroscience 38*, 29 (2018), 6461–6474.

[15] Friston, K., Ashburner, J., Kiebel, S., Nichols, T., and Penny, W. *Statistical Parametric Mapping (SPM12)*. Wellcome Trust Centre for Neuroimaging, London, UK, 2014. Available at: http://www.fil.ion.ucl.ac.uk/spm/.

[16] Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., and Frackowiak, R. S. J. Spatial registration and normalization of images. *Human Brain Mapping 3*, 3 (1995), 165–189.

[17] Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., and Frackowiak, R. S. J. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping 2*, 4 (1994), 189–210.

[18] Gousias, I. S., Rueckert, D., Heckemann, R. A., Dyet, L. E., Boardman, J. P., Edwards, A. D., and Hammers, A. Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. *NeuroImage 40*, 2 (2008), 672–684.

[19] Görgen, K., Hebart, M. N., Allefeld, C., and Haynes, J.-D. The same analysis approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. *NeuroImage 180* (2018), 19–30. New advances in encoding and decoding of brain signals.

[20] Haynes, J.-D., and Rees, G. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience 7* (Jul 2006), 523–534.

[21] Hebart, M. N., Görgen, K., and Haynes, J.-D. The decoding toolbox (tdt): A versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics 8* (2015).

[22] Henson, R. N. A., Buechel, C., Josephs, O., and Friston, K. J. The slice-timing problem in event-related fmri. In *Proceedings of the 5th International Conference on Functional Mapping of the Human Brain* (Düsseldorf, Germany, June 1999).

[23] Horing, B., and Büchel, C. The human insula processes both modality-independent and pain-selective learning signals. *PLOS Biology 20*, 5 (05 2022).

[24] Koenen, L. R., Pawlik, R. J., Icenhour, A., et al. Associative learning and extinction of conditioned threat predictors across sensory modalities. *Communications Biology 4* (2021), 553.

[25] Kriegeskorte, N., and Bandettini, P. Analyzing for information, not activation, to exploit high-resolution fmri. *NeuroImage 38*, 4 (Dec 2007), 649–662. Epub 2007 Feb 27.

[26] Kriegeskorte, N., Goebel, R., and Bandettini, P. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences 103*, 10 (2006), 3863–3868.

[27] Krishnan, A., Woo, C.-W., Chang, L. J., Ruzic, L., Gu, X., López-Solà, M., Jackson, P. L., Pujol, J., Fan, J., and Wager, T. D. Somatic and vicarious pain are represented by dissociable multivariate brain patterns. *eLife 5* (2016), e15166.

[28] Lindquist, M. A. The statistical analysis of fmri data. *Statistical Science 23*, 4 (2008), 439–464.

[29] Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging 16*, 2 (April 1997), 187–198.

[30] Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., and Mourão-Miranda, J. Quantitative prediction of subjective pain intensity from whole-brain fmri data using gaussian processes. *NeuroImage 49*, 3 (2010), 2178–2189.

[31] Mitchell, S. D., and O'Sullivan, M. *PuLP: A Linear Programming Toolkit for Python*. The University of Auckland Operations Research Group, 2011.

[32] Mouraux, A., and Iannetti, G. D. Nociceptive laser-evoked brain potentials do not reflect nociceptive-specific neural activity. *Journal of Neurophysiology 101*, 6 (2009), 3258–3269.

[33] Mumford, J., Turner, B., Ashby, F., and Poldrack, R. Deconvolving bold activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage 59*, 3 (2 2012), 2636–2643.

[34] Nemhauser, G., and Wolsey, L. *Integer and Combinatorial Optimization*. John Wiley & Sons, New York, 1988.

[35] Nichols, T. E., and Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping 15*, 1 (Jan 2002), 1–25.

[36] Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. Beyond mind-reading: Multi-voxel pattern analysis of fmri data. *Trends in Cognitive Sciences 10*, 9 (September 2006), 424–430.

[37] Palermo, S., Benedetti, F., Costa, T., and Amanzio, M. Pain anticipation: an activation likelihood estimation meta-analysis of brain imaging studies. *Human Brain Mapping 36*, 5 (May 2015), 1648–1661.

[38] Pearce, J. M., and Hall, G. A model for pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review 87*, 6 (1980), 532–552.

[39] Poldrack, R. A., Mumford, J. A., and Nichols, T. E. Preprocessing fmri data. In *Handbook of Functional MRI Data Analysis*. Cambridge University Press, Cambridge, 2011, pp. 34–52.

[40] Rescorla, R. A., and Wagner, A. R. *A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement.* Appleton-Century-Crofts, New York, NY, 1972, pp. 64–99.

[41] Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., and Greicius, M. D. Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience 27*, 9 (2007), 2349–2356.

[42] Sladky, R., Friston, K., Tröstl, J., Cunnington, R., Moser, E., and Windischberger, C. Slice-timing effects and their correction in functional mri. *NeuroImage 58*, 2 (Sep 2011), 588–594. Epub 2011 Jul 2.

[43] Todd, M. T., Nystrom, L. E., and Cohen, J. D. Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage 77* (2013), 157–165.

[44] Turner, B. O., Mumford, J. A., Poldrack, R. A., and Ashby, F. G. Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage 62*, 3 (Sep 2012), 1429–1438.

[45] Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage 180* (2018), 68–77.

[46] Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage 145* (2017), 166–179. Individual Subject Prediction.

[47] Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., and Kross, E. An fmri-based neurologic signature of physical pain. *New England Journal of Medicine 368*, 15 (2013), 1388–1397.

[48] Wild, H. M., Heckemann, R. A., Studholme, C., and Hammers, A. Gyri of the human parietal lobe: Volumes, spatial extents, automatic labelling, and probabilistic atlases. *PLOS ONE 12*, 8 (08 2017), 1–28.

[49] Worsley, K. J., Liao, C.-H., Aston, J., Petre, V., Duncan, G. H., Morales, F., and Evans, A. C. A general statistical analysis for fmri data. *NeuroImage 15*, 1 (Jan 2002), 1–15.

A Appendix

A.1 Explanation of Preprocessing Steps

A.1.1 Slice Timing Correction

Slice timing correction is an essential preprocessing step in fMRI data analysis. Because fMRI data are typically collected in a sequential manner—either ascending, descending, or interleaved—slices are acquired at different times during each repetition time (TR) period. These temporal offsets can introduce inaccuracies when interpreting the timing of neural events and the corresponding blood oxygen level-dependent (BOLD)[1] signal changes, which are crucial for event-related designs and studies requiring precise temporal resolution [7, 22, 42].

The goal of slice timing correction is to adjust the time series of each voxel so that the acquired data reflect the assumption that all slices were obtained simultaneously. This adjustment is achieved by interpolating the magnetic resonance (MR) signal for each slice to a common reference time point within the TR period, often the midpoint. Various interpolation methods, such as spline, Fourier, or polynomial interpolation, can be used depending on the data characteristics and analysis requirements [7, 42, 22].

A.1.2 Realignment

Realignment is critical in correcting head movements during scanning. Even minor head movements can introduce significant noise into fMRI data, confounding the results by altering the spatial consistency of the data.

[1] BOLD signal measures brain activity by detecting changes in blood oxygenation. When neurons become active, they use more oxygen, causing increased blood flow to that area. The MRI scanner detects these blood flow changes, allowing researchers to identify active brain regions.

In this work, realignment was conducted in two steps. First, each volume was aligned to the initial volume of the time series to correct for movements between image acquisitions. Then, all volumes were aligned to the mean image derived from the volumes, ensuring more accurate correction over the entire session [16]. The movement parameters obtained from this process (translations and rotations) can also be used as nuisance regressors in statistical analyses to further reduce the impact of residual movement effects [2].

A.1.3 Co-registration

Co-registration aligns the mean functional image to the participant's high-resolution anatomical T1 image. Functional images typically have lower resolution and different contrast properties compared to anatomical images. Co-registration adjusts for differences in image orientation, scaling, and resolution.

Linear transformations correct for translations and rotations between the functional and anatomical images, while nonlinear adjustments allow for finer corrections to match the underlying brain structures. The resulting alignment ensures that functional data can be accurately localized to anatomical regions of the brain for further analysis [2, 10, 29].

A.1.4 ROI Transformation and Thresholding

To accurately align the insula ROIs with each subject's native anatomical space, inverse deformation fields obtained during the normalization process in the preprocessing pipeline were employed. The insula ROI masks were derived from probabilistic atlases in Montreal Neurological Institute (MNI) space, with voxel-wise probability values ranging from 0 to 30. These values indicate the consistency of voxel inclusion within the insula across subjects in the atlas: a value of 30 signifies that all subjects included that voxel within the respective insula region, while a value of 15 indicates that half of the subjects had that voxel within the region.

Transformation Process:

1. **Inverse Deformation Fields**: Inverse deformation fields were applied to transform the MNI space probabilistic insula maps back to each subject's native anatomical space. This transformation accounts for individual anatomical variability, ensuring that the ROIs correspond accurately to each subject's brain anatomy.

2. **Thresholding**: After transformation, the probability maps retained their original value range (0 to 30). To create binary ROIs, a threshold of 15 was applied. Voxels with probability values ≥ 15 were included in the ROI, indicating that these voxels were present in at least 50% of the subjects within the atlas. This threshold strikes a balance between anatomical specificity and inclusivity, ensuring that the ROIs are both consistent across subjects and representative of the insula regions of interest.

3. **Interpolation Method**: Nearest-neighbor interpolation was used during the transformation process to preserve the discrete nature of the region labels in the probability maps. This method minimizes the introduction of partial voluming effects, maintaining the integrity of the ROI boundaries.

4. **Validation**: Visual inspections were conducted for a subset of subjects to ensure proper alignment with anatomical landmarks. Additionally, overlap metrics were calculated to assess the consistency of ROI placement across subjects.

A.2 Understanding Least Squares Separate (LSS)

In fMRI analysis, it is often desired to understand how the brain reacts to specific events, such as seeing an image or feeling a sensation. However, brain data are complex, and signals from multiple trials can overlap. To address this, a method called Least Squares Separate is used. This technique helps separate the brain's response to each individual trial, providing clearer insights into how the brain processes different events over time [33, 44].

LSS works within a statistical framework called the General Linear Model (GLM), which is applied to analyze how different experimental events contribute to the observed brain signals [17, 28]. The basic form of the GLM is:

$$Y = X \cdot \beta + \epsilon \tag{A.1}$$

Where:

- Y is the observed fMRI data (brain activity over time),
- X is the design matrix, which represents the timing and type of experimental events (e.g., when a stimulus was presented),

- β represents the strength of the brain's response to each event,
- ϵ is the residual error (the part of the signal that is not explained) [39, 49].

In traditional fMRI analysis, multiple trials are often averaged together. LSS, however, assigns a separate regressor to each individual trial, providing clearer insights into brain activity patterns on a trial-by-trial basis [33]. The model for a single trial can be written as:

$$Y = X_{\text{trial}} \cdot \beta_{\text{trial}} + X_{\text{nuisance}} \cdot \beta_{\text{nuisance}} + \epsilon \tag{A.2}$$

Where X_{trial} is the design matrix for the trial of interest, β_{trial} represents the response to that trial, and X_{nuisance} includes all other trials grouped together as a nuisance regressor [33]. This method allows isolation of the neural response for each specific event without interference from other trials.

LSS is particularly useful in decoding studies, such as MVPA, where detailed trial-specific information is crucial for training machine learning models to decode complex neural patterns [20, 25, 36].

A.3 Mathematical Formulation of the ILP Optimization

Variables and Sets

- Let $H = \{1, 2, \ldots, N_h\}$ denote the set of heat trials.
- Let $S = \{1, 2, \ldots, N_s\}$ denote the set of sound trials.
- Let $F = \{1, 2, \ldots, N_f\}$ represent the set of CV folds.

Decision Variables

The binary decision variable x_{ijk} indicates whether heat trial $i \in H$ is paired with sound trial $j \in S$ within fold $k \in F$:

$$x_{ijk} = \begin{cases} 1 & \text{if heat trial } i \text{ is paired with sound trial } j \text{ in fold } k, \\ 0 & \text{otherwise.} \end{cases}$$

Objective Function

The objective was to minimize the sum of absolute differences in trial numbers between paired heat and sound trials, thereby balancing the signal across folds:

$$\min \sum_{k \in F} \sum_{i \in H} \sum_{j \in S} c_{ij} x_{ijk},$$

where $c_{ij} = |\text{Trial}_i - \text{Trial}_j|$ represents the cost associated with pairing trials based on their trial numbers.

Constraints

1. **Pairing Constraints:**
   - Each heat trial can be paired with at most one sound trial within each fold:
   $$\sum_{j \in S} x_{ijk} \leq 1, \quad \forall i \in H, \, \forall k \in F$$
   - Each sound trial can be paired with at most one heat trial within each fold:
   $$\sum_{i \in H} x_{ijk} \leq 1, \quad \forall j \in S, \, \forall k \in F$$
2. **Fold Size Constraint:** Each fold is required to contain exactly P pairs of heat and sound trials:
   $$\sum_{i \in H} \sum_{j \in S} x_{ijk} = P, \quad \forall k \in F$$

3. **CS Index Balance Constraint:** Paired heat and sound trials must have balanced values of the CS Index within each fold:

$$\sum_{i \in H} \sum_{j \in S} x_{ijk} \cdot \text{CSID}_i = \sum_{i \in H} \sum_{j \in S} x_{ijk} \cdot \text{CSID}_j, \quad \forall k \in F$$

4. **Session Number Balance Constraint:** The session numbers between paired trials must be balanced:

$$\sum_{i \in H} \sum_{j \in S} x_{ijk} \cdot \text{Session}_i = \sum_{i \in H} \sum_{j \in S} x_{ijk} \cdot \text{Session}_j, \quad \forall k \in F$$

5. **Previous Modality Balance Constraint:** Paired heat and sound trials must have balanced values of the previous modality measure within each fold:

$$\sum_{i \in H} \sum_{j \in S} x_{ijk} \cdot \text{PrevMod}_i = \sum_{i \in H} \sum_{j \in S} x_{ijk} \cdot \text{PrevMod}_j, \quad \forall k \in F.$$

6. **Uniqueness Constraint:** Each trial can be used at most once across all folds:

$$\sum_{k \in F} \sum_{j \in S} x_{ijk} \leq 1, \quad \forall i \in H$$

$$\sum_{k \in F} \sum_{i \in H} x_{ijk} \leq 1, \quad \forall j \in S$$

Note: Not all balance constraints (CS Index, Session Number, and Previous Modality) are necessarily applied in every implementation.

A.4 Manhattan Distance Definition and Illustrative Example

A.4.1 Definition

The Manhattan distance D is used to measure deviations from an ideal balance of trial number distributions. For this analysis, the temporal order of trials is represented as

indices corresponding to each label (e.g., heat and sound). Given a sequence of n trials, let the indices for one label (e.g., heat) be

$$\mathbf{P} = \big[P_1, P_2, \ldots, P_{n/2}\big],$$

and the indices for the other label (e.g., Sound) be

$$\mathbf{Q} = \big[Q_1, Q_2, \ldots, Q_{n/2}\big].$$

The Manhattan distance is then computed as:

$$D = \sum_{i=1}^{n/2} \big|P_i - Q_i\big|. \tag{A.3}$$

This metric captures the total deviation between the indices of trials belonging to different labels. Smaller values of D indicate more interleaving between heat and sound trials, whereas larger values indicate clustering.

A.4.2 Ideal Interleaving Case

In the ideal case, where the indices of trials for each label are perfectly interleaved, the Manhattan distance reaches its minimal value. Consider a sequence with n trials, evenly split between two labels. The optimal distance D_{optimal} can be derived as follows:

$$D_{\text{optimal}} = \sum_{i=1}^{n/2} \big|(2i-1) - (2i)\big| = \sum_{i=1}^{n/2} 1 = \frac{n}{2}. \tag{A.4}$$

A.4.3 Example

To better illustrate, consider a sequence of six trials, with two labels: A (heat) and B (sound).

Perfectly Balanced Sequence

$$\text{Ideal sequence: } A, B, A, B, A, B, \tag{A.5}$$

with the indices represented as:

$$\mathbf{P} = [1, 3, 5], \quad \mathbf{Q} = [2, 4, 6]. \tag{A.6}$$

Hence,

$$D_{\text{optimal}} = |1 - 2| + |3 - 4| + |5 - 6| = 1 + 1 + 1 = 3. \tag{A.7}$$

Clustered Sequence

However, if the observed sequence clusters trials of one label, such as:

$$\text{Observed sequence: } A, A, B, A, B, B, \tag{A.8}$$

then the indices become:

$$\mathbf{P} = [1, 2, 4], \quad \mathbf{Q} = [3, 5, 6]. \tag{A.9}$$

The Manhattan distance is:

$$D = |1 - 3| + |2 - 5| + |4 - 6| = 2 + 3 + 2 = 7, \tag{A.10}$$

which is substantially higher than the ideal $D_{\text{optimal}} = 3$.

This indicates that label A (heat) appears clustered together, rather than interleaved with label B (sound).

A.5 Prevalence Inference Explanation

The aim of the algorithm introduced by Allefeld et al (2016) [1] is to estimate whether an information effect (e.g., above-chance accuracy) is present in a significant proportion of the population for each ROI or voxel, while controlling for multiple comparisons across ROIs. Both the global null hypothesis (no effect in any subject) and prevalence null hypothesis (effect is present in at most γ_0 fraction of the population) are tested.

Notation and Inputs

- N: Number of subjects.

- r: A region of interest.
- real_accuracies$[r, k]$: Accuracy of subject k for ROI r using the original (unpermuted) labels.
- permuted_accuracies$[r, k, i]$: Accuracy of subject k for ROI r using the i-th permutation of labels. Each subject has P_1 possible permutations, including the real data as $i = 1$.
- P_2: Number of second-level permutations (randomly combining first-level permutations across subjects).
- α: Significance level (e.g., 0.05).
- γ_0: Threshold prevalence to test (e.g., 0.5 to test whether a majority of subjects show the effect).

Algorithm Steps

Step 1: Compute the Real Data Minimum.

For each ROI r, the minimum accuracy across subjects is computed using the real (unpermuted) data:

$$m_r = \min_{k=1,\ldots,N} \big(\text{real\_accuracies}[r, k]\big). \tag{A.11}$$

This step identifies the lowest classification performance among all subjects for each ROI under the observed data.

Step 2: Generate Second-Level Permutations.

To establish a null distribution, P_2 second-level permutations are created by randomly selecting one permutation index for each subject. Let

$$\text{perms}[k]$$

denote the permutation index chosen for subject k. For a second-level permutation $j \in \{1, \ldots, P_2\}$:

- For $j = 1$: The original data are used for all subjects, i.e.,

$$\text{perms}[k] = 1 \quad \forall k.$$

- For $j > 1$: A random permutation index is selected for each subject, i.e.,

$$\text{perms}[k] \in \{1, 2, \ldots, P_1\}.$$

This step ensures a random combination of label permutations across subjects.

Step 3: Compute the Minimum Statistic for Each Permutation.

For each second-level permutation j and ROI r, the minimum accuracy across subjects is calculated as:

$$m_{r,j} = \min_{k=1,\ldots,N} \big(\text{permuted\_accuracies}[r, k, \text{perms}[k]]\big). \tag{A.12}$$

The minimum statistic $m_{r,j}$ represents the lowest classification accuracy in ROI r for the given permutation j.

Step 4: Count Permutations Greater Than or Equal to the Real Minimum.

The number of permutations where the permuted minimum $m_{r,j}$ is greater than or equal to the real-data minimum m_r is counted. For each ROI r, this count is updated as:

$$\text{uRank}[r] \leftarrow \text{uRank}[r] + \begin{cases} 1, & \text{if } m_{r,j} \geq m_r, \\ 0, & \text{otherwise.} \end{cases} \tag{A.13}$$

Step 5: Compute the Uncorrected p-Value for the Global Null.

After all P_2 permutations, the uncorrected p-value for the global null hypothesis (no effect in any subject) is calculated as:

$$p_N(m_r) = \frac{\text{uRank}[r]}{P_2}. \tag{A.14}$$

This value reflects the proportion of second-level permutations where the minimum statistic is as large or larger than the observed minimum.

Step 6: Correct for Multiple Comparisons Using the Maximum Statistic.

To control for multiple comparisons across all ROIs, a maximum statistic is computed for each second-level permutation:

$$M_j = \max_r \big(m_{r,j}\big). \tag{A.15}$$

The familywise error rate (FWER) corrected p-value for the global null at ROI r is then:

$$p_N^*(m_r) = \frac{1}{P_2} \sum_{j=1}^{P_2} \Big[M_j \geq m_r \Big]. \tag{A.16}$$

This step identifies how often the maximum minimum statistic across ROIs exceeds the observed minimum for ROI r.

Step 7: Test for Prevalence Above a Threshold γ_0.

The prevalence null hypothesis tests whether the effect is present in at most γ_0 fraction of the population. It is shown that:

$$p_N(m_r \mid \gamma \leq \gamma_0) \leq \Big[(1-\gamma_0) \sqrt[N]{p_N(m_r)} + \gamma_0 \Big]^N. \tag{A.17}$$

Using this, an adjusted significance level for prevalence inference is computed:

$$\alpha_r^* = \frac{\alpha - p_N^*(m_r)}{1 - p_N^*(m_r)}. \tag{A.18}$$

From this, a lower bound $\gamma_{0,r}$ on the prevalence can be derived:

$$\gamma_{0,r} = \frac{{\alpha_r^*}^{\frac{1}{N}} - p_N(m_r)^{\frac{1}{N}}}{1 - p_N(m_r)^{\frac{1}{N}}}. \tag{A.19}$$

Summary of the Algorithm

1. **Within each subject**:

- Compute classification accuracies for real data (permutation $i = 1$).
- Compute classification accuracies for all label permutations ($i = 2, \ldots, P_1$).

2. **Across subjects**:
   - For each ROI r, compute the real-data minimum m_r.
   - Perform P_2 second-level permutations:
     a) If $j = 1$, use all real data.
     b) If $j > 1$, randomly select one permutation index perms$[k] \in \{1, \ldots, P_1\}$ for each subject k and compute $m_{r,j}$.
   - Count permutations where $m_{r,j} \geq m_r$ to compute $p_N(m_r)$.
   - Use the maximum statistic across ROIs to calculate the corrected $p_N^*(m_r)$.
   - Derive prevalence bounds $\gamma_{0,r}$ based on α and $p_N(m_r)$.

This approach provides valid population-level inferences, delivering corrected p-values for the presence of effects and lower bounds on the fraction of the population showing above-chance performance in each ROI.

Example: 3 Subjects, 2 ROIs, and 4 Permutations

This example demonstrates how to apply all steps of the prevalence-inference algorithm, including multiple-comparisons correction and prevalence inference. Consider 3 subjects ($N = 3$), 2 ROIs, and 4 permutations per subject ($P_1 = 4$).

Real Data (Permutation 1)

$$\text{real\_accuracies} = \begin{bmatrix} 70\%, & 65\%, & 60\% \\ 80\%, & 85\%, & 78\% \end{bmatrix}.$$

The columns denote subjects, the rows denote ROIs (ROI 1 on the first row, ROI 2 on the second row).

Permuted Data

$$\text{permuted\_accuracies} = \left[\begin{bmatrix} 70 & 68 & 66 & 64 \\ 65 & 63 & 61 & 67 \\ 60 & 62 & 64 & 66 \end{bmatrix}, \quad \begin{bmatrix} 80 & 79 & 77 & 76 \\ 85 & 83 & 82 & 86 \\ 78 & 81 & 79 & 80 \end{bmatrix} \right].$$

Each column corresponds to one of the $P_1 = 4$ first-level permutations for each subject ($i = 1, \ldots, 4$), with column 1 being the real/unpermuted data.

Step 1: Compute the Real Data Minimum for Each ROI

$$m_{\text{real, ROI 1}} = \min(70\%, 65\%, 60\%) = 60\%,$$

$$m_{\text{real, ROI 2}} = \min(80\%, 85\%, 78\%) = 78\%.$$

Step 2: Generate Second-Level Permutations

Each second-level permutation j is formed by choosing one of the four first-level permutations for each subject. For example:

$$j = 1 : [1, 1, 1], \quad j = 2 : [2, 3, 4], \quad j = 3 : [3, 2, 4], \quad j = 4 : [4, 4, 2].$$

Step 3: Compute the Minimum Statistic $m_{r,j}$ for Each ROI and Each Second-Level Permutation

$$m_{r,j} = \min_{k=1,\ldots,N} \Big(\text{permuted\_accuracies}[r,\, k,\, \text{perms}[k]] \Big).$$

Concretely, for each j:

1. Permutation 1 (real data): $[1, 1, 1]$

$$\text{ROI 1} : [70\%, 65\%, 60\%] \;\rightarrow\; m_{1,\text{ROI 1}} = 60\%,$$

$$\text{ROI 2} : [80\%, 85\%, 78\%] \;\rightarrow\; m_{1,\text{ROI 2}} = 78\%.$$

2. Permutation 2: $[2, 3, 4]$

$$\text{ROI 1} : [68\%,\ 61\%,\ 66\%] \ \rightarrow\ m_{2,\text{ROI 1}} = 61\%,$$
$$\text{ROI 2} : [79\%,\ 82\%,\ 80\%] \ \rightarrow\ m_{2,\text{ROI 2}} = 79\%.$$

3. Permutation 3: $[3, 2, 4]$

$$\text{ROI 1} : [66\%,\ 63\%,\ 66\%] \ \rightarrow\ m_{3,\text{ROI 1}} = 63\%,$$
$$\text{ROI 2} : [77\%,\ 83\%,\ 80\%] \ \rightarrow\ m_{3,\text{ROI 2}} = 77\%.$$

4. Permutation 4: $[4, 4, 2]$

$$\text{ROI 1} : [64\%,\ 67\%,\ 62\%] \ \rightarrow\ m_{4,\text{ROI 1}} = 62\%,$$
$$\text{ROI 2} : [76\%,\ 86\%,\ 81\%] \ \rightarrow\ m_{4,\text{ROI 2}} = 76\%.$$

Step 4: Count Permutations Meeting or Exceeding the Real Minimum

For ROI 1: $(m_{1,\text{ROI 1}}, m_{2,\text{ROI 1}}, m_{3,\text{ROI 1}}, m_{4,\text{ROI 1}}) = (60\%, 61\%, 63\%, 62\%)$.

For ROI 2: $(m_{1,\text{ROI 2}}, m_{2,\text{ROI 2}}, m_{3,\text{ROI 2}}, m_{4,\text{ROI 2}}) = (78\%, 79\%, 77\%, 76\%)$.

The real-data minimum is $m_{1,r}$. Hence, for ROI 1, $m_{\text{real, ROI 1}} = 60\%$. All four permutation values $\{60\%, 61\%, 63\%, 62\%\}$ are $\geq 60\%$, so

$$\text{uRank[ROI 1]} = 4.$$

For ROI 2, $m_{\text{real, ROI 2}} = 78\%$. Among $\{78\%, 79\%, 77\%, 76\%\}$, two of the values (78% and 79%) are $\geq 78\%$. So

$$\text{uRank[ROI 2]} = 2.$$

Step 5: Uncorrected *p*-Values for the Global Null

$$p_N\big(m_{\text{ROI 1}}\big) = \frac{\text{uRank[ROI 1]}}{P_2} = \frac{4}{4} = 1.0,$$
$$p_N\big(m_{\text{ROI 2}}\big) = \frac{\text{uRank[ROI 2]}}{P_2} = \frac{2}{4} = 0.5.$$

Step 6: Correct for Multiple Comparisons Using the Maximum Statistic

The maximum statistic across ROIs in each permutation is computed as

$$M_j = \max_{r\in\{\text{ROI 1, ROI 2}\}}\big(m_{r,j}\big).$$

Hence, for each of the 4 permutations:

$$M_1 = \max(60\%,\, 78\%) = 78\%,$$
$$M_2 = \max(61\%,\, 79\%) = 79\%,$$
$$M_3 = \max(63\%,\, 77\%) = 77\%,$$
$$M_4 = \max(62\%,\, 76\%) = 76\%.$$

To compute the FWER–corrected p-value for ROI 1 with observed minimum $m_{1,\text{ROI 1}} = 60\%$, it is determined how many of the M_j exceed or equal 60%:

$$[M_1,\, M_2,\, M_3,\, M_4] = [78\%,\, 79\%,\, 77\%,\, 76\%]$$

are all $\geq 60\%$. Thus

$$p_N^*(m_{\text{ROI 1}}) = \frac{4}{4} = 1.0.$$

Similarly, for ROI 2 with observed minimum $m_{1,\text{ROI 2}} = 78\%$, the number of $M_j \geq 78\%$ is:

$$\{78\%, 79\%, 77\%, 76\%\} \quad \rightarrow \quad 2 \text{ values} \geq 78\%.$$

Hence:

$$p_N^*(m_{\text{ROI 2}}) = \frac{2}{4} = 0.5.$$

Step 7: Prevalence Inference

The prevalence null hypothesis tests whether the fraction of subjects γ with real effects is $\leq \gamma_0$. For illustration, assume $\gamma_0 = 0.5$. The relevant formulae (see the main text) can now be applied. In the preceding example, ROI 2 yielded:

1. An FWER-corrected p-value of $p_N^*(m_r) = 0.50$. Using $\alpha = 0.05$,

$$\alpha_r^* = \frac{\alpha - p_N^*(m_r)}{1 - p_N^*(m_r)} = \frac{0.05 - 0.50}{1 - 0.50} = \frac{-0.45}{0.50} = -0.90.$$

   Since α_r^* is negative, it indicates that there is insufficient evidence to reject the global or prevalence null at the usual $\alpha = 0.05$ level for ROI 2.

2. If α_r^* had been positive, the lower bound on effect prevalence, $\gamma_{0,r}$, could have been computed via

$$\gamma_{0,r} = \frac{{\alpha_r^*}^{\frac{1}{N}} - p_N(m_r)^{\frac{1}{N}}}{1 - p_N(m_r)^{\frac{1}{N}}}.$$

Consider now a new ROI 3 with a more extreme observed minimum. Suppose it yields

$$p_N(m_r) = 0.002 \quad \text{and} \quad p_N^*(m_r) = 0.04.$$

Repeating the same procedure:

1. The adjusted significance level becomes

$$\alpha_r^* = \frac{0.05 - 0.04}{1 - 0.04} = \frac{0.01}{0.96} \approx 0.0104,$$

   which is now positive.

2. For $N = 3$, the lower bound on effect prevalence is

$$\gamma_{0,r} = \frac{{\alpha_r^*}^{\frac{1}{3}} - (0.002)^{\frac{1}{3}}}{1 - (0.002)^{\frac{1}{3}}}.$$

   If ${\alpha_r^*}^{\frac{1}{3}} \approx 0.215$ and $(0.002)^{\frac{1}{3}} \approx 0.126$, then

$$\gamma_{0,r} \approx \frac{0.215 - 0.126}{1 - 0.126} = \frac{0.089}{0.874} \approx 0.102.$$

Thus it can be concluded that at least 10.2% of the population exhibits above-chance accuracy in ROI 3 at $\alpha = 0.05$, indicating a positive lower bound on prevalence.

This hypothetical scenario shows how a stronger effect (smaller p-values) can yield a positive α_r^* and thereby a nonzero $\gamma_{0,r}$, providing evidence that a certain fraction of the population truly has above-chance performance.

Interpretation and Summary of the Example

- **ROI 1**: The minimum in real data (60%) is so low that all permutations produce minima at least as high. Consequently, the uncorrected and corrected p-values are both 1.0, which does not support rejecting the null hypothesis for this ROI.
- **ROI 2**: The uncorrected p-value is 0.5. After correcting for multiple comparisons, it remains 0.5. In this artificial example, the observed minimum (78%) does not appear extreme enough to reject the null or infer a high prevalence of effects in ROI 2 at $\alpha = 0.05$.
- **ROI 3 (Hypothetical)**: In contrast, if a more extreme observed minimum yielded smaller p-values (e.g., an FWER-corrected p^* below 0.05), a positive adjusted significance level α_r^* could be obtained, supporting rejection of the prevalence null for at least some fraction of the population.

This example illustrates the complete chain of computations:

1. Within each subject, accuracies for real and permuted data are obtained.
2. Across subjects, minima are computed for each second-level permutation and compared to the real-data minimum to form an uncorrected null distribution.
3. A familywise error rate correction is then carried out using the maximum statistic across ROIs.
4. Finally, the logic of prevalence inference is applied to determine whether an effect is present in a significant fraction of the population.

Although the prevalence null is not rejected in the first two ROIs of this toy scenario, the procedure demonstrates how both global null testing and prevalence bounds can be derived in a unified framework. In real data analyses, stronger observed effects typically lead to smaller p-values and a positive lower bound on the prevalence of the information effect (as shown for the hypothetical ROI 3).

Also, because only $P_2 = 4$ second-level permutations are considered, p-values can only take values in multiples of $1/4 = 0.25$. Hence, it is impossible to achieve $p < 0.05$ with such a small permutation count. In real applications, more permutations (e.g., hundreds or thousands) are typically performed, allowing much finer resolution of p-values and the possibility of detecting significant results below the 0.05 threshold.

A.6 Searchlight Mapping and Voxel-Wise One-Sample t-Tests

A searchlight analysis was performed to examine which local brain regions reliably distinguish between heat and sound trials. This method evaluates classification accuracy in small, spherical neighborhoods of voxels across the entire brain, rather than restricting the analysis to predefined regions of interest [25, 26]. As illustrated in Figure A.1, a sphere with a radius of 10 mm is iteratively centered on each voxel in the brain, and a classifier is trained and tested on the data from that local neighborhood.

At each step of the procedure, the resulting classification accuracy is assigned to the central voxel, producing an "accuracy map" for every participant. These accuracy maps reflect how well the local voxel cluster around each position can discriminate the two conditions (heat vs. sound). As with the ROI-based analyses, single-trial betamaps were used, and CV folds were optimized to mitigate systematic confounds.

Once the accuracy maps were generated, they were normalized to the standard Montreal Neurological Institute (MNI) space using SPM, enabling voxel-wise comparisons across participants. The normalized accuracy maps were then submitted to voxel-wise one-sample t-tests against the theoretical chance level of 50%. Two main contrasts were assessed:

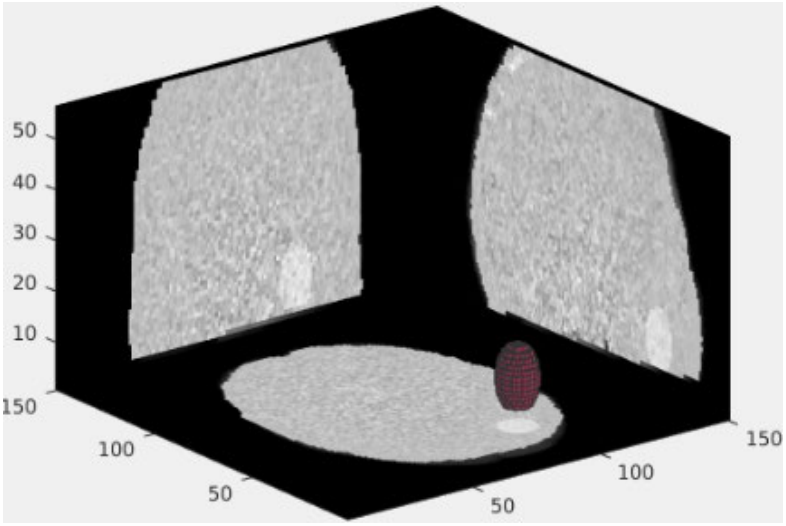

Figure A.1: **Searchlight Mapping Visualization.** This figure illustrates the searchlight mapping technique, where a spherical region (searchlight) is systematically moved across the brain to identify locally informative voxel patterns. The radius of the searchlight was set to 10 mm to balance spatial resolution and computational efficiency.

1. **Positive Contrast:** Tested whether classification accuracy was significantly above 50%, indicating that local neural patterns in that region reliably discriminated the two aversive modalities.

2. **Negative Contrast:** Tested whether classification accuracy was significantly below 50%, which may suggest the presence of systematic misclassification or confounding factors in that region.

FWE correction was applied to manage the increased risk of false positives due to the large number of voxel-wise tests [35, 49]. Clusters surviving FWE correction were interpreted as locations where local patterns of brain activity effectively encoded modality-specific information.

Compared with the ROI-based approach, the whole-brain searchlight analysis allows detection of smaller-scale or spatially unexpected patterns of discrimination. Accordingly, any significant findings beyond the primary ROIs point to additional brain areas that may also contribute to encoding the difference between aversive modalities.

A.7 Second Level Results Uncontrolled

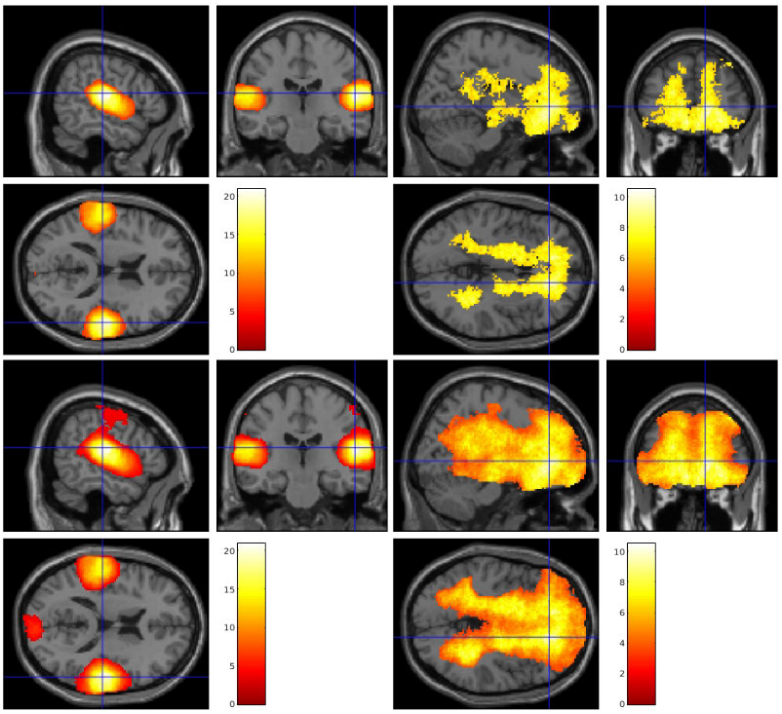

Figure A.2: **Actual Modality: Second-Level Results of Uncontrolled CV Designs.** Heatmaps of t-values for positive (left) and negative (right) contrasts from t-tests. The upper row shows results corrected for FWE, and the lower row shows uncorrected results ($p < 0.001$). The respective crosshair is positioned at the global maximum activation. Brighter colors (red and yellow) represent higher t-values. For the positive contrast, two significant clusters can be observed, while for the negative contrast, a substantial number of significant voxels are evident.

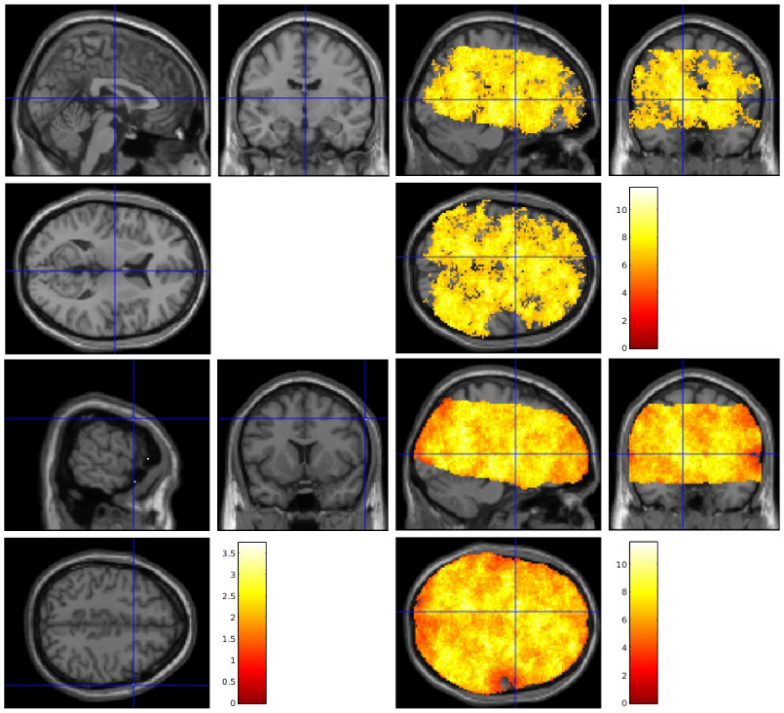

Figure A.3: **Expected Modality: Second-Level Results of Uncontrolled CV Designs.** Heatmaps of t-values for positive (left) and negative (right) contrasts from t-tests. The upper row shows results corrected for FWE, and the lower row shows uncorrected results ($p < 0.001$). The crosshair is positioned at the respective global maximum activation, if present. Brighter colors (red and yellow) represent higher t-values. No significant voxels were observed for the positive contrast under FWE correction, with only a few single voxels appearing when uncorrected. Conversely, a substantial number of significant voxels were observed for the negative contrast.

A.8 Tools and Software Used

Table A.1 lists the tools and software packages used in the preparation of this master's thesis.

Table A.1: Tools and Software Used

Tool	**Purpose**
SPM12	Functional MRI data preprocessing and analysis
TDT	Multivariate pattern analysis
Python, MATLAB	Data processing and analysis scripting
PuLP	Integer Linear Programming for cross-validation
LaTeX	Thesis preparation and document typesetting system
Claude, ChatGPT	Technical writing and documentation assistance

Erklärung zur selbständigen Bearbeitung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort	Datum	Unterschrift im Original