# Stimulus-Driven Leakage in Naturalistic Neuroimaging

Seung-Goo Kim*

Research Group Neurocognition of Music and Language,

Max Planck Institute for Empirical Aesthetics, Grüneburgweg 14, 60322 Frankfurt, Germany

*Correspondence: seung-goo.kim@ae.mpg.de

April 14, 2026

**Keywords:** encoding analysis, predictive modelling, data leakage, cross-validation, overfitting

**Abstract:** This article elucidates a methodological pitfall of cross-validation for evaluating predictive models applied to naturalistic neuroimaging data——namely, "stimulus-driven leakage." While this problem has been well known as "leakage in training examples" in machine learning, it may be difficult to detect in practice due to conventions in neuroscience. Stimulus-driven leakage can occur when predictive modelling is applied to data from a conventional neuroscientific design, characterized by a limited set of stimuli repeated across trials and/or participants. It results in spurious predictive performance due to overfitting to repeated signals, even in the presence of independent noise. Through comprehensive simulations and real-world examples, following a theoretical formulation, the article underscores how such data leakage can occur and how severely it can compromise results and conclusions when combined with widely spread informal reverse inference. The article concludes with practical recommendations for researchers to avoid stimulus-driven leakage in their experimental design and analysis.

# 1 Introduction

## 1.1 Data leakage in predictive modelling

In statistical learning and machine learning, evaluation of a learned model is a critical step since the model is, by design, highly flexible in learning patterns in the data, which simultaneously gives rise to the possibility of overfitting to noise that always exists in the data. Thus, it is important to evaluate the model's performance on an independent test set that was not used during the training and optimisation processes. Surprisingly often, designing a legitimate evaluation can be non-trivial (Kaufman et al., 2012). The key aspect is to keep the training, validation (i.e., optimisation), and test partitions completely separated. When this fails, data may 'leak' into the model from the validation or test set, which must be kept 'sealed'. The model may then behave over-optimistically, leading to wrong conclusions such as overconfidence in its generalisability and overestimation of feature importance. This is known as 'data leakage', and it remains a major challenge in predictive modelling, even in recent machine-learning-based scientific research (Kapoor & Narayanan, 2023), including neuroimaging (Rosenblatt et al., 2024; Verstynen & Kording, 2023).

## 1.2 Data leakage in naturalistic neuroimaging

While the predictive modelling has been heavily used in clinical neuroimaging research (e.g., brain age gap estimation; Seitz-Holland et al., 2024), it remains relatively novel in many cognitive research domains. This paper primarily focuses on a specific field, often called 'naturalistic neuroimaging'. It refers to the use of complex, real-world stimuli (e.g., movies, music, natural speech) in neuroimaging experiments to investigate brain function in more ecologically valid contexts (Hamilton & Huth, 2020; Nastase et al., 2020; Sonkusare et al., 2019). In particular, this idea has been widely adopted in domains where high-order cognitive and/or affective processes are involved, and a simple contrastive experimental approach can explain only little. For example, comparing brain responses to *music* vs. *non-music* to find neural correlates of "music perception" may follow an overly reductionist assumption (i.e., "music-as-fixed-effect" fallacy; Kim, 2022) that the human brain is governed by simple rules that can be extrapolated to explain complex behaviours (for more discussion, see Nastase et al., 2020).

Two different approaches have been popularised in studies addressing the complexity of real-world information. First, as a model-based approach, a temporal transfer functions are estimation, following the electrophysiological tradition of the receptive field mapping (Lalor et al., 2009; Theunissen et al., 2000; Wu et al., 2006). Second, as a model-free approach, a intersubject (or intertrial) correlation is used to

quantify the strength of the stimulus-driven signal in the data (Hasson et al., 2004).

A problem arises when the model-based approach is applied to a dataset collected for the model-free approach without due caution. In the model-free approach, the repeated stimuli across trials and/or participants are essential to identify stimulus-driven effects. However, in the model-based approach, the same stimulus repeated across CV partitions constitutes data leakage, even if the noise is independent across partitions. In this paper, I refer to this specific form of leakage as *Stimulus-driven Leakage* (SDL) since the repeated stimulus is the source of the leakage. Importantly, SDL is not inherent to the model-based approach or to the use of naturalistic stimuli itself, but arises from their improper combination. According to Kaufman et al., 2012, SDL can be understood as a special form of *leakage in training examples*, which can also be seen as *Non-independence between train and test samples* in a recently proposed taxonomy (Kapoor & Narayanan, 2023).

## 1.3 Aims and scope of the current paper

While the true prevalence of the stimulus-driven leakage in published studies has yet to be found, a preliminary investigation suggests that it is not uncommon in the literature of naturalistic neuroscience including fMRI and M/EEG studies. In particular, while the accessibility of the predictive modelling has largely increased thanks to various open-source software packages (e.g., `mTRF Toolbox` [Crosse et al., 2016]; `scikit-learn` [Pedregosa et al., 2011]), some researchers may not be fully aware of the pitfalls of the predictive modelling, despite efforts to educate researchers about the best practices in encoding analysis (Crosse et al., 2021; Dupré la Tour et al., 2025). This overall situation increases the risk of the data leakage, which could lead to contamination of the literature. Therefore, this paper aims to explain the mechanism of the stimulus-driven leakage and to demonstrate how it can occur and how severely it can compromise the results.

The current paper focuses on a particular form of predictive modelling: a time series prediction with a finite-impulse response model that fits variable neural delays, regularised with a ridge penalty. This form of predictive modelling is widely used in the naturalistic neuroimaging, especially when involving temporally dynamic stimuli (e.g., movies, music, natural speech). Nonetheless, stimulus-driven leakage can occur in any model in which the same stimuli exist in both training and test sets.

For those who are trained in machine learning (or statistical learning), this may seem obvious and intuitive. However, it may not be so clear to classically trained neuroscientists who are unfamiliar with machine learning methods. In the Theory section, a formal analysis shows how the intuition that the repeated stimulus constitutes data leakage stands. In the Simulations section, contributing factors that worsen

the stimulus-driven leakage are identified. The Real-Data section demonstrates the effects of SDL based on real-world datasets. The Discussion section discusses the implications of SDL for future analyses and experiments, and provides practical recommendations to avoid it.

# 2 Theory

We are interested in predicting a time series response from a set of features using a linear model. Consider a linear model:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^{T \times 1}$ is a response vector over $T$ time points of a single response unit (e.g., a channel or a voxel), $\mathbf{X} \in \mathbb{R}^{T \times FD}$ is a Toeplitz design matrix constructed from $F$ features[1] (e.g., the audio envelope) with $D$ delays, forming a finite impulse response (FIR) model, $\mathbf{b} \in \mathbb{R}^{FD \times 1}$ is an unknown weight vector, $\mathbf{e} \in \mathbb{R}^{T \times 1}$ is a zero-mean, unit-variance Gaussian noise vector $\mathbf{e} \sim \mathcal{N}_T(\mathbf{0}, \mathbf{I}_T)$ where $\mathbf{I}_T \in \mathbb{R}^{T \times T}$ is an identity matrix. For convenience, features (i.e., before Toeplitz construction) and response variables are assumed to be standardized prior to analysis so that their sample means are zero and their variances are one. Please note that this paper focuses on the FIR model for its popularity in time-invariant linear system identification (Crosse et al., 2021; Huth et al., 2016; Wu et al., 2006). Nonetheless, the conclusion of this section generalises to any design matrix (i.e., $\mathbf{X}$ does not need to be a Toeplitz matrix).

A ridge solution (Hoerl & Kennard, 1970) to Equation 1 is given by

$$\hat{\mathbf{b}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda \mathbf{I}_{FD}\right)^{-1} \mathbf{X}^\mathsf{T}\mathbf{y}, \tag{2}$$

where $\lambda > 0$ is a regularisation hyperparameter.

In cross-validation (CV), the datasets are partitioned into three independent sets: training, validation, and test sets. The training set is used to estimate the weights $\hat{\mathbf{b}}$ for a given $\lambda$, the validation set is used to find the optimal $\lambda$ that maximises the (validation) prediction accuracy, and the test set is used to estimate the out-of-sample prediction accuracy with the optimal $\lambda$.

For a true model ($\mathbf{X}$) with a sufficient signal strength ($\|\mathbf{X}\mathbf{b}\|_2^2 \gg 0$), the prediction accuracy (either Pearson correlation or $R^2$) with a minimal regularisation ($\lambda \to 0$) is expected to be positive whereas that of a null model (i.e., reasonable random features $\mathbf{U}$) with a proper regularisation ($\lambda \gg 0$) is expected

[1] While *feature* and *predictor* are often interchangeably used, in this article a feature refers to a variable that describes the characteristics of interest of the input, while a predictor refers to a feature with a specific delay (i.e., each column of a design matrix). Thus, $F$ features with $D$ delays produce $P = FD$ predictors.

to be null. This is how the CV is supposed to works in usual cases. However, when the same stimulus is repeated across CV partitions, the expected prediction accuracy of the null model can be positive.

But in which cases can the same stimulus be repeated across partitions? To illustrate, consider two possible modelling approaches for a dataset in which two stimuli were presented to three subjects (Figure 1). One is subject-specific modelling (Figure 1a; e.g., leave-one-stimulus-out CV) where each subject's data are partitioned into training, validation, and test sets. The other is stimulus-specific modelling (Figure 1b; e.g., leave-one-subject-out CV). In the stimulus-specific modelling, the same stimulus is repeated across CV partitions, albeit with different noise realizations, since the same stimulus is presented to all subjects.

While it may seem obvious to those trained in machine learning that the latter is a flawed CV design, I argue that stimulus-specific modelling may appear legitimate to researchers from other disciplines. This apparent legitimacy may be due to the independence of noise across partitions, or simply because of the familiarity with leave-one-*subject*-out CV in other analyses (e.g., multi-voxel pattern analysis [MVPA]; Kriegeskorte et al., 2006). In particular, avoiding the repetition of identical noise in analyses has been established as a common practice in neuroimaging, largely thanks to the pedagogical work that coined the term "double-dipping" (Kriegeskorte et al., 2009). Where double-dipping concerns the repetition of identical noise, SDL concerns the repetition of identical signal. Given that the data is the sum of signal and noise, SDL can be thought of as "inverse double-dipping".

The mechanism of SDL can be summarized in the following steps:

1. The repeated signal disables the seemingly legitimate regularisation process.
2. As the regularisation hyperparameter (i.e., $\lambda$) approaches zero, the projection matrix becomes positive definite, which would have been null if properly regularised.
3. As a result, the expected prediction accuracy of the null model over random realisations is proportional to a bilinear form involving a non-zero vector and a positive-definite square matrix, which is positive.

Below are the brief explanations of the above points. A more detailed derivation is given in the Supplementary Theory.

## 2.1 Disabling regularisation

To denote partitions in cross-validation, let us use the subscript $i$ for the $i$-th partition: $\mathbf{X}_1$ and $\mathbf{y}_1$ are the training set, $\mathbf{X}_2$ and $\mathbf{y}_2$ the validation set to optimise training, and $\mathbf{X}_3$ and $\mathbf{y}_3$ the test set to estimate

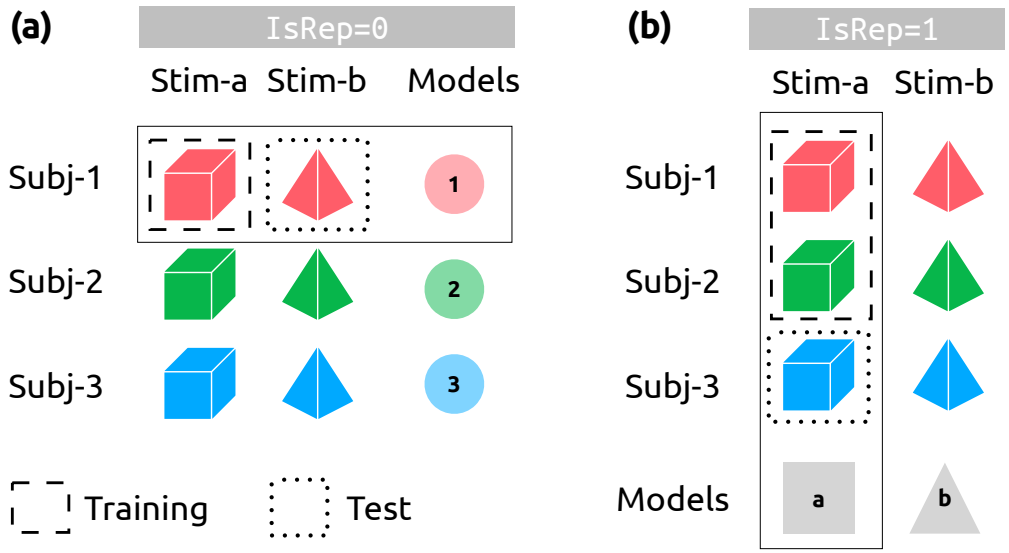


Figure 1: Schema of two cross-validation designs. For simplicity, let us assume that we have three subjects $\{1, 2, 3\}$ (depicted in red, green, blue) and two stimuli $\{a, b\}$ (depicted as a cube and a tetrahedron). Note that we assume any repetitions of stimuli within a subject are averaged prior to the cross-validation. **(a)** `IsRep=0`. Subject-specific models (pale coloured circles). Each model (e.g., pale red circle in solid rectangle) is trained with one stimulus (dashed rectangle) and tested on another stimulus (dotted rectangle). **(b)** `IsRep=1`. Stimulus-specific models (gray square and triangle). Each model (e.g., gray square in solid rectangle) is trained with two subjects (dashed rectangle) and tested on the other subject (dotted rectangle). The validation set, which could be in the training set of the outer loop, is not marked for simplicity. The CV design (b) suffers from SDL while the CV design (a) does not.

out-sample prediction performance. When the identical stimulus is used in all sets ($\mathbf{s}_1 = \mathbf{s}_2 = \mathbf{s}_3$), the expected validation accuracy (i.e., Pearson correlation) of the null model $\mathbf{U}$ across random noise realisations is proportional to the following bilinear form:

$$\mathbb{E}\left[\text{corr}\left(\hat{\mathbf{y}}_2[\mathbf{U}],\, \mathbf{y}_2\right)\right] \propto \mathbf{s}_1^{\mathsf{T}} \mathbf{P}_{\mathbf{U}}^{\mathsf{T}} \mathbf{s}_1, \tag{3}$$

where the projection matrix based on the null model is $\mathbf{P}_{\mathbf{U}} = \mathbf{U}_1 \left(\mathbf{U}_1^{\mathsf{T}}\mathbf{U}_1 + \lambda \mathbf{I}\right)^{-1} \mathbf{U}_1^{\mathsf{T}}$.

In this case, the optimal $\lambda$ that maximises the validation can be analytically found using singular vector decomposition (Hastie et al., 2009, Eq. 3.47):

$$\lambda^* = \arg\max_{\lambda} \mathbf{s}_1^{\mathsf{T}} \left( \sum_{j=1}^{P} \upsilon_{(j)} \frac{\delta_j^2}{\delta_j^2 + \lambda} \upsilon_{(j)}^{\mathsf{T}} \right) \mathbf{s}_1 = \varepsilon \approx 0, \tag{4}$$

where $\upsilon_{(j)}$ is a left singular vector of $\mathbf{U}$, $\delta_j$ is the corresponding singular value, and $\varepsilon$ is the possible smallest positive value (i.e., the regulaisation is effectively disabled).

## 2.2 Positive definite projection matrix

When unregularised (i.e., $\lambda^* \approx 0$), the projection matrix based on the null model can be approximated as a positive definite matrix:

$$\mathbf{P}_{\mathbf{U}} = \mathbf{U}_1 \left(\mathbf{U}_1^{\mathsf{T}}\mathbf{U}_1 + \lambda^* \mathbf{I}\right)^{-1} \mathbf{U}_1^{\mathsf{T}} \overset{(4)}{\approx} \mathbf{U}_1 \left(\mathbf{U}_1^{\mathsf{T}}\mathbf{U}_1\right)^{-1} \mathbf{U}_1^{\mathsf{T}}, \tag{5}$$

This is not the case with proper regularisation (i.e., $\lambda^* \gg 0$) where the projection matrix is $\mathbf{P}_{\mathbf{U}} \approx \frac{1}{\lambda^*}\mathbf{U}_2\mathbf{U}_1^{\mathsf{T}} \approx \mathbf{0}$. (Equation S13 in the Supplementary Theory).

## 2.3 Positive null prediction accuracy

Therefore, the expected out-sample (i.e., test) accuracy of the null model is:

$$\mathbb{E}\left[\text{corr}\left(\hat{\mathbf{y}}_3[\mathbf{U}],\, \mathbf{y}_3\right)\right] \propto \mathbf{s}_1^{\mathsf{T}} \mathbf{P}_{\mathbf{U}}^{\mathsf{T}} \mathbf{s}_1 \overset{(5)}{\approx} \mathbf{s}_1^{\mathsf{T}} \mathbf{U}_1 \left(\mathbf{U}_1^{\mathsf{T}}\mathbf{U}_1\right)^{-1} \mathbf{U}_1^{\mathsf{T}} \mathbf{s}_1 > 0. \tag{6}$$

$\mathbf{U}_1^\mathsf{T}\mathbf{U}_1$ is positive definite due to its symmetry and independence of the columns in $\mathbf{U}_1$. Since the inverse operation preserves the signs of eigenvalues of a square matrix, its inversion $\left(\mathbf{U}_1^\mathsf{T}\mathbf{U}_1\right)^{-1}$ is also positive definite. Thus, the expected prediction accuracy (i.e., a bilinear form involving a non-zero vector and a positive-definite square matrix) of the null model is positive.

Equation 6 clearly suggests that any random features may seem to predict 'unseen' (but leaked in truth) data. Depending on the signal-to-noise ratio (SNR), this may result in significant Type-I (false positive) errors.

# 3 Simulations

For a graphical illustration, a simple case of small-scale simulation (i.e., a toy example) is shown in Figure 2. In this example, 100 time points for 3 response variates were generated with two features and three delays. The SNR was 0 dB. In the first case (Figure 2a), the stimulus was not repeated across CV partitions. Thus, as expected, the prediction accuracies for the null models (pink) were around zero and the optimal ridge penalties were large (i.e., $\lambda \gg 0$). However, in the second case (Figure 2b), the stimulus was repeated across CV partitions. Because of the stimulus-driven leakage, the prediction accuracies for the null models were well-above the threshold corresponding to Bonferroni-adjusted one-tailed $P$-value of 0.05, and the optimal ridge penalties were similar to that of the true models. Extended figures with more detailed explanations can be found in the Supplementary Results (Figure S1–2).

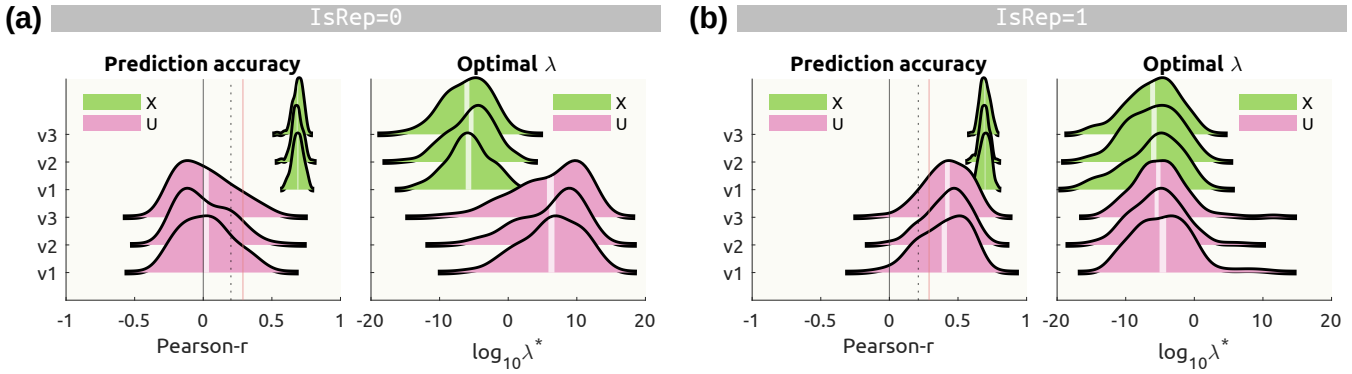


Figure 2: A toy example of stimulus-driven leakage. For each case when (a) the stimulus is not repeated across CV partitions and (b) the stimulus is repeated across CV partitions, prediction accuracies (left) based on the true features (lime green) and null features (pink) over 200 simulations are shown with the optimal ridge penalties (right). White vertical bands of each ridgeline represent the 95% confidence interval of the mean. Vertical lines for Pearson correlation mark non-parametrically estimated thresholds corresponding to one-tailed $P = 0.05$ (gray dashed), and its Bonferroni adjustment (red solid).

The spurious inflation of the prediction accuracy due to the stimulus-driven leakage (i.e., the SDL artefact) is positively proportional to the SNR.

When exploring the contributing factors to the strength of the SDL artefact, it was revealed that the SDL artefact is greater when the null model is more flexible (i.e., with more features, with less autocorrelation, and more delays Figure 3a,b,c) and the true features have strong autocorrelation (Figure 3d). See the Supplementary Results for the extensive simulations quantifying the effect sizes of these factors.

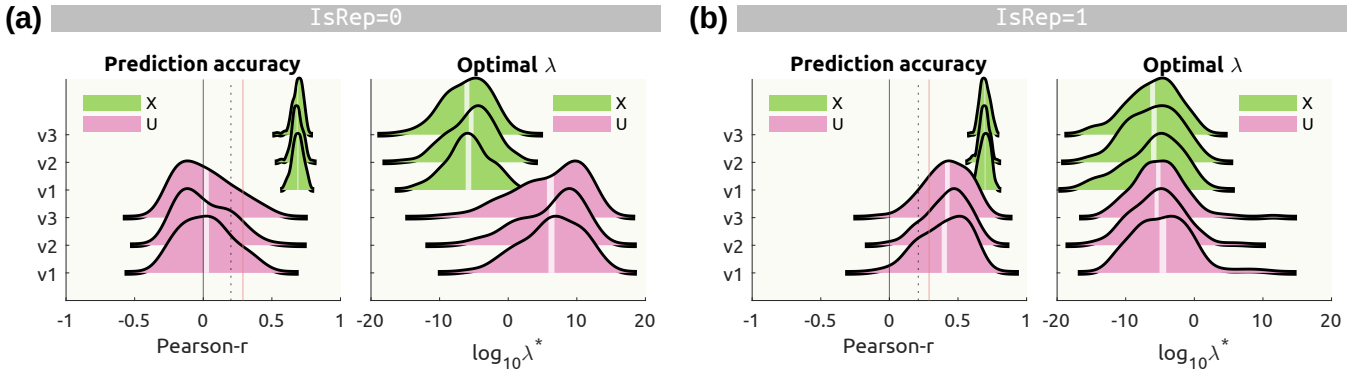


Figure 3: PLACEHOLDER

# 4 Real Data

Having explained the theory and shown simulations, a logical empirical question would be—*Could SDL happen with real data in a significant manner?* In this section, I demonstrate real-data examples of the SDL artefact using open-access data where healthy participants listened to various musical excerpts while measuring neural activity (electroencephalography [EEG] or functional magnetic resonance imaging [fMRI]) or behavioural ratings (Kaneshiro et al., 2020; Sachs et al., 2020). For clarification, none of these studies reported results that suffer from SDL. The datasets were used for their classical design (i.e., the same stimuli set for all participants) to demonstrate a fictitious analysis that introduces the SDL.

True features were the audio envelopes extracted from the musical stimuli using a cochlear model (Chi et al., 2005). This is based on the well-established encoding of the acoustic energy in the human auditory system (Ding & Simon, 2013). Null features were either (a) the phase-randomised envelope (preserving spectral magnitudes and autocorrelation structures) as the most realistic null feature, (b) normal noise, and (c) uniform noise as the least realistic null feature. Details of the real data and analysis implementation are provided in the Supplementary Materials.

The magnitude of the SDL artefact was estimated as the difference in null prediction accuracies between two CV schemes (i.e., leave-$N$-stimulus-out for subject-specific modelling vs. leave-$N$-subject-out for stimulus-specific modelling): $\mathrm{SDL} = \bar{r}_{\mathrm{stim}}(\mathbf{U};\, \mathtt{IsRep} = 1) - \bar{r}_{\mathrm{subj}}(\mathbf{U};\, \mathtt{IsRep} = 0)$ where $r_{\mathrm{stim}}$ is the prediction accuracy of a stimulus-specific model and $r_{\mathrm{subj}}$ is that of a subject-specific model. $\bar{(\cdot)}$ denotes averaging across null models. If there were no false inflation of prediction accuracy due to SDL, the null predictions with and without stimulus repetitions should be equal (i.e., $\mathcal{H}_0 : \mathbb{E}(\mathrm{SDL}) = 0$). Otherwise, the null prediction with stimulus repetitions is expected to be greater than the null prediction without repetitions ($\mathcal{H}_A : \mathbb{E}(\mathrm{SDL}) > 0$).

All data analyses were consistently done using the MATLAB package Linearised Encoding Analysis (LEA; https://github.com/seunggookim/lea).

## 4.1 Electroencephalography

Scalp electrical potential data were recorded in 48 healthy participants while listening to Western-style Indian pop music (i.e., Bollywood music; Kaneshiro et al., 2020). Using this EEG dataset, linearised encoding analysis was performed with an assumed 'true' feature (i.e., the audio envelope), the three types of null features (phase-randomised envelopes, normal noise, and uniform noise), and various delay ranges (0–0.3 sec, 0–0.5 sec, and 0–1 sec).

Figure 4 displays the results of the encoding analysis with the phase-randomised envelope as the null feature and the delay range of 0–0.5 sec (64 samples) as an example. Without stimulus repetition ($\mathtt{IsRep} = 0$), a clear fronto-central topography is shown in the prediction accuracy ($\max r(X;0) = 0.047$, Figure 4**a**) as well as in the ridge hyperparameter ($\min \log_{10} \lambda(X;0) = 5.86$, Figure 4**b**), reflecting the envelope encoding in the bilateral auditory cortices while listening to music. For this particular data, the estimated weights were stronger in the left than right fronto-central channels (Figure 4**c**). With the phase-randomised envelope, as expected, the null prediction accuracy was minimal ($\max r(U;0) = 0.006$, Figure 4**d**; $\max \mathbb{E}\left[r(U;0)\right] = 0.001$, Figure 4**g**) with all channels being highly regularised ($\min \log_{10} \lambda(U;0) = 10.98$, Figure 4**e**; $\min \mathbb{E}\left[\log_{10} \lambda(U;0)\right] = 12.11$, Figure 4**h**).

With stimulus repetition ($\mathtt{IsRep} = 1$), true prediction accuracies were increased ($\max r(X;1) = 0.085$, Figure 4**j**). This is because, unlike the simulations where we knew the true feature, our 'true' feature (i.e., the audio envelope) was not the sole information that the human EEG data encode. That is, the EEG data may encode other features of the stimulus (e.g., timbre, rhythm, pitch) that are not modelled, and thus the prediction accuracy can be inflated by the stimulus repetition.

However, most strikingly, the null prediction accuracies with stimulus repetition showed an almost identical topography to the actual encoding results (Figure 4**m,p**), with even higher values than the true prediction without repetition ($\max r(X;0) = 0.047$, $\max r(U;1) = 0.052$, $\max \mathbb{E}\left[r(U;1)\right] = 0.056$). Note that, by definition, the null feature (phase-randomised envelopes) should not have predicted anything in the EEG data. However, when the identical stimuli were repeated over CV partitions, the regularisation was disabled—regardless of the given features—in channels where the stimulus-evoked response is strong (Figure 4**n,q**). Then, even random weights (Figure 4**o**; see Supplementary Figure S14 for time series), which were widely different from the true weights, could successfully predict the repeated signal. Because the SDL artefact reflects the genuine biological signal that is driven by the repeated stimulus, the observed pattern of the prediction accuracy is indistinguishable from the true signal (unless investigating the weights; see also see Supplementary Figures S14–16).

Figure 5 shows the SDL effects with the phase-randomised envelope, normal noise, and uniform noise. The SDL effects were found to be significant in most channels except for the frontopolar electrodes ($P_{FDR} < 0.01$), exhibiting a fronto-central topography that is commonly found in association with the auditory cortical activity. It is noteworthy that the SDL effect is much stronger for the phase-randomised envelope, which preserves the autocorrelation structure of the stimulus. This implies that the SDL artefact is more severe when the null features have strong autocorrelation, which can be the case for any arbitrary features that are extracted from naturalistic stimuli (e.g., movies, music, natural speech). Also, the number of delays seems to further exacerbate the SDL artefact, consistent with the simulation results.

## 4.2 Functional magnetic resonance imaging

Blood-oxygen-level-dependent (BOLD) data were acquired in 39 healthy participants while listening to Western instrumental musical pieces that either evoke happiness or sadness as validated in independent listeners (Sachs et al., 2020). As done for the EEG dataset, linearised encoding analysis was performed with the fMRI data as responses, the audio envelope as a true feature, the phase-randomised envelope as a null feature, and delays from 3 to 9 seconds (7 samples) (Figure 6). Similarly to the EEG results, the phase-randomised envelope strikingly predicted the BOLD time series in the bilateral auditory cortices including the Heschl's gyrus and planum temporale (Figure 6**m,p**) while no consistent pattern in the transfer function weights was found over the phase randomizations (Figure 6**r**), clearly demonstrating the SDL effect. Once again, the anatomical location and the extent of the heightened null prediction accuracies precisely matched the true encoding results (Figure 6**a,j**), which would seem 'highly convincing' to many neuroscientists.

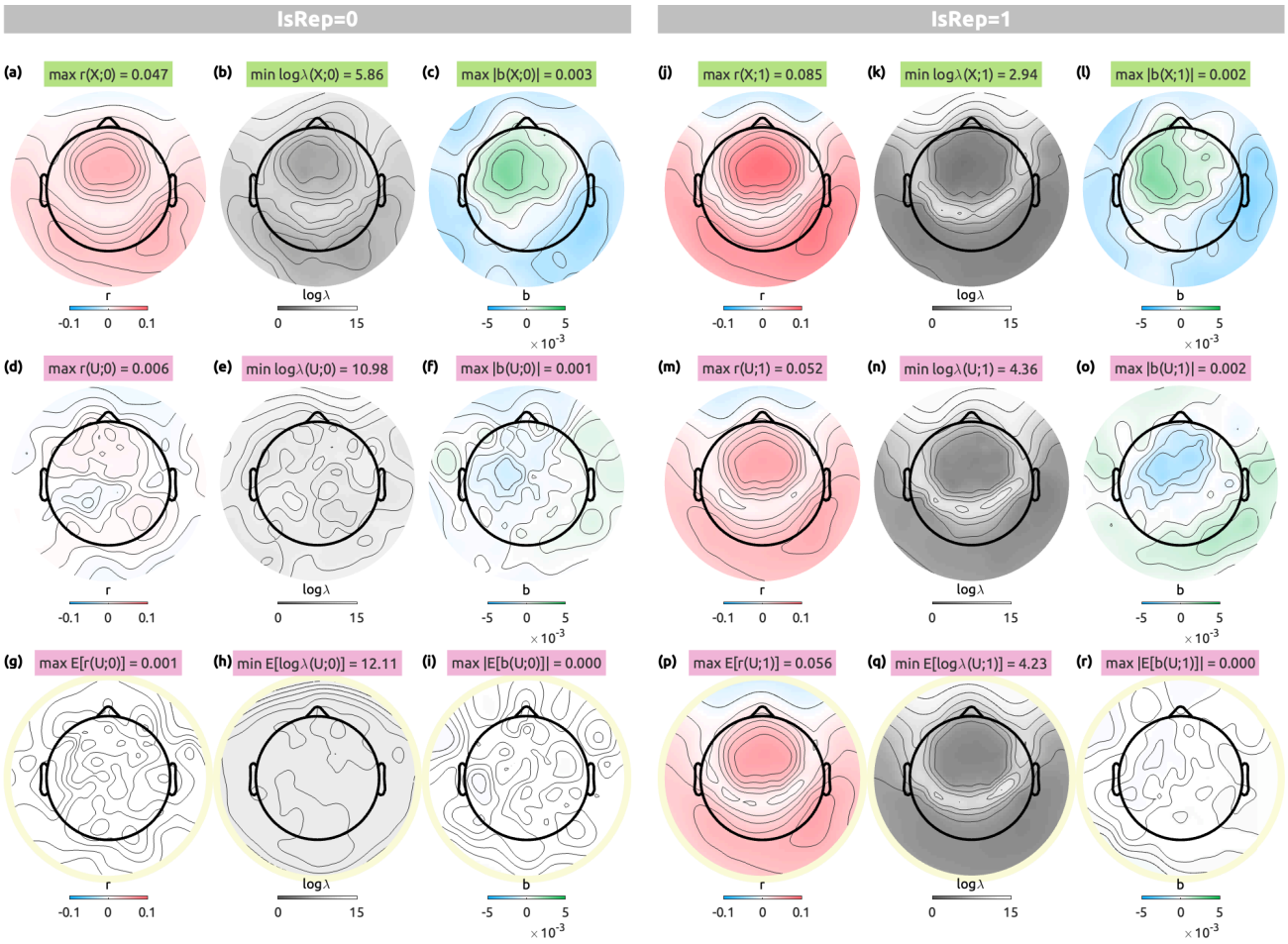


Figure 4: EEG linearised encoding analysis results with delays from 0 to 0.5 sec with an audio envelope (top row, **(a-c, j-l)**), a single case of a phase-randomised envelope (middle row, **(d-f, m-o)**), and an average of 100 phase-randomised envelopes (bottom row, circled in pale yellow, **(g-i, p-r)**). For each CV scheme ($\texttt{IsRep} = 0$, left panels, **(a-i)**; $\texttt{IsRep} = 1$, right panels, **(j-r)**), prediction accuracy ($r$, blue to red, **(a, d, g, j, m, p)**), logarithmic ridge hyperparameter ($\log_{10} \lambda$, gray to white, **(b, e, h, k, n, q)**), transfer function weights that are summed over delays ($b$, blue to green, **(c, f, i, l, o, r)**) are shown along the columns. Note that stimulus repetition not only slightly inflated true prediction accuracies but even the predicted 'brain activity pattern' from null features (**m,p**), which was effectively indistinguishable from true predictions (**a, j**).

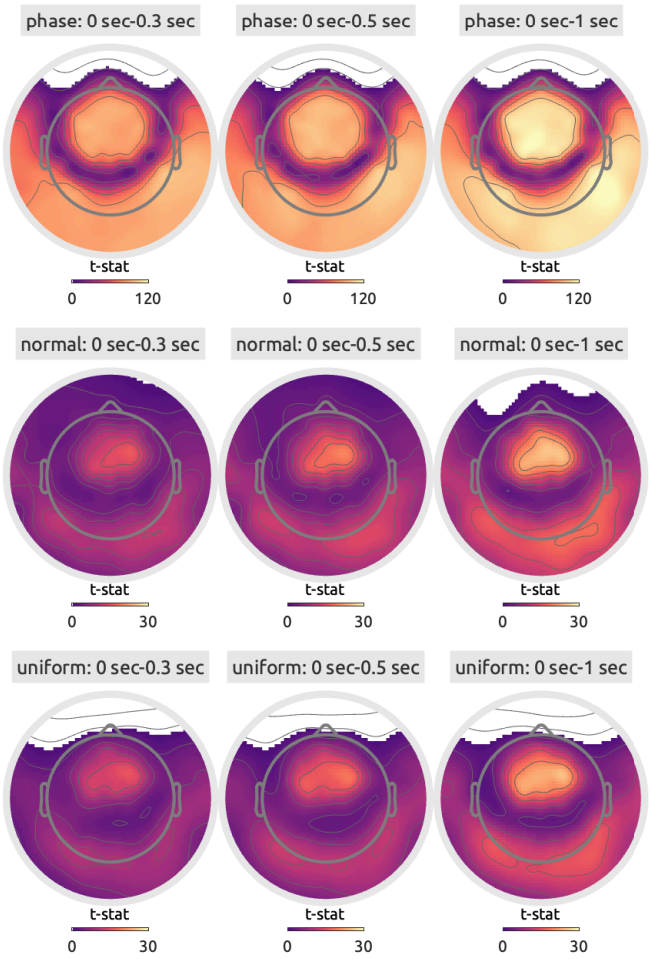


Figure 5: $t$-statistic maps comparing the null prediction accuracies between two CV schemes (e.g., Figure 4**g** vs. Figure 4**p**) to test the SDL effects in the EEG data for the phase-randomised envelope (top row), the normal noise (middle row), and the uniform noise (bottom row). Channels were thresholded by statistical significance with FDR adjustment (one-sided $P_{FDR} < 0.01$).

When analysed with different noise models and delays (see Supplementary Figures S17–26), the similar SDL pattern was consistently observed (i.e., inflated prediction accuracies in the auditory cortices). The SDL effect was statistically significant not only in the bilateral superior temporal gyri but also in the medial occipital cortices and the inferior frontal cortices, where acoustic energy is not expected to be encoded (Figure 7). Consistent with the EEG results, the SDL effect was stronger for the phase-randomised envelope than the normal or uniform noise as well as for more delays than less ones.

The weights (Supplementary Figures S27–29) display prominent "auditory components" even for normal and uniform noise in their eigenvectors while the corresponding eigenvariates were widely different from the weights estimated by the true feature.

## 4.3 Behavioural ratings

Continuous ratings of music-evoked emotions were sampled from the same 39 healthy participants who took part in the fMRI experiment above (Sachs et al., 2020). After the scanning session, participants listened to the same musical pieces again and rated their Emotionality (how happy/sad they felt) and Enjoyment (how much they enjoyed the piece) using a slider. Figure 8 shows the results of the encoding analysis with the phase-randomised envelope as the null feature and the delay range of 0–10 sec (51 samples) as an example. Once more, while the true envelope predicted Emotionality to some degree and Enjoyment to a greater extent (Figure 8**a**), the phase-randomised envelope also predicted both scales well above zero when the stimuli were repeated across CV partitions (Figure 8**m,p**) unlike when the stimuli were not repeated (Figure 8**g**).

The SDL effect was significant also in the behavioural ratings consistently across all noise models and delays (Figure 9; see also Supplementary Figures 20–37). Similarly to other modalities, the transfer function weights reflected the inherent autocorrelation structure of the behavioural data (Supplementary Figures 38–40).

# 5 Discussion

The primary objective of cognitive neuroscience is to comprehend how the brain executes information-processing operations (Kay, 2018). Linearised encoding analysis serves as a robust method to evaluate a model (i.e., transfer function) that describes how the brain encodes sensory information from the environment and processes this information further (Naselaris et al., 2011). However, it is critical to

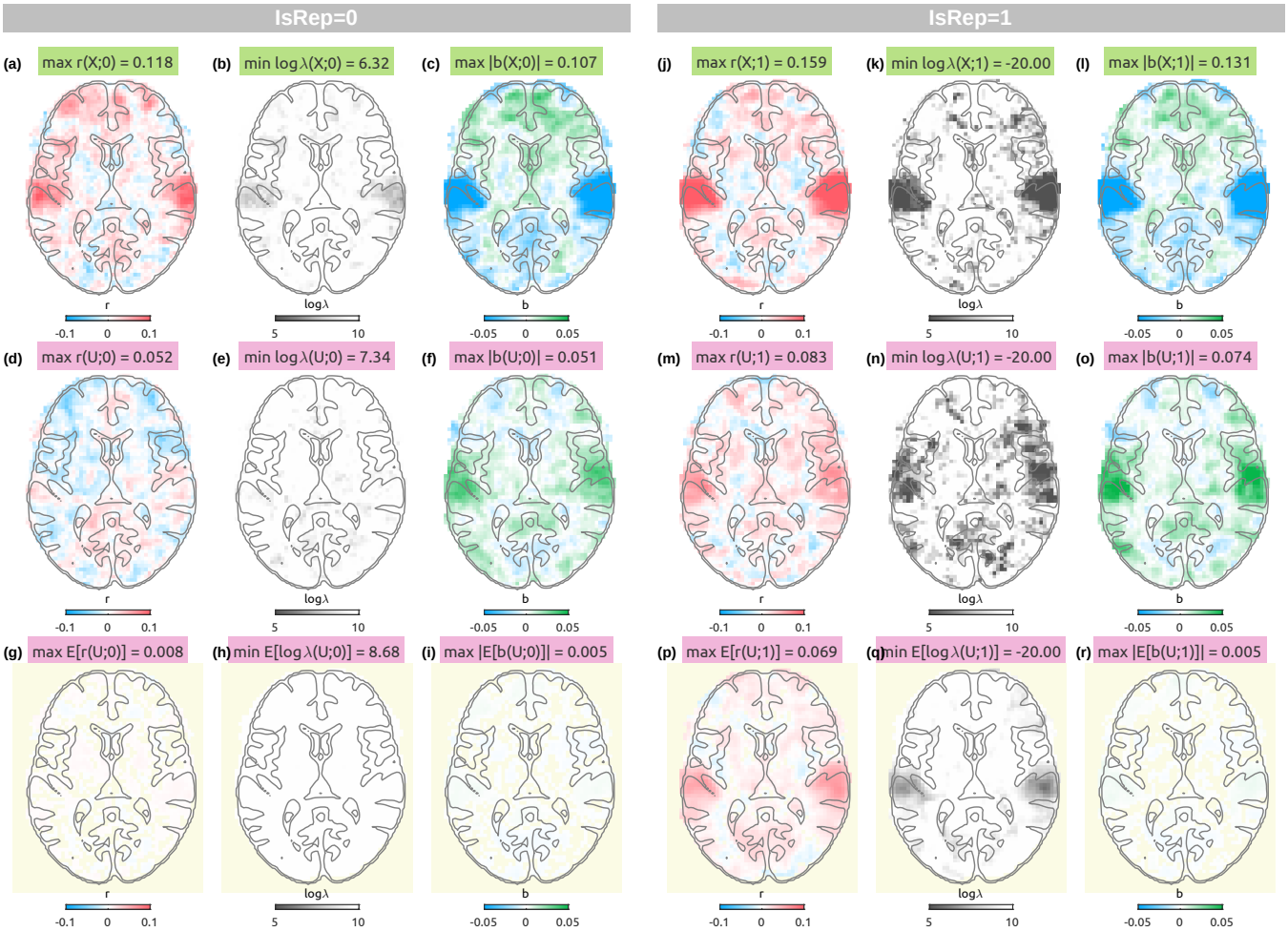


Figure 6: fMRI linearised encoding analysis results with delays from 3 to 9 sec with the audio envelope (top row, **(a-c, j-l)**), a single case of a phase-randomised envelope (middle row, **(d-f, m-o)**), and an average of 100 phase-randomised envelopes (bottom row, pale yellow background, **(g-i, p-r)**). For each CV scheme ($\texttt{IsRep} = 0$, left panels, **(a-i)**; $\texttt{IsRep} = 1$, right panels, **(j-r)**), prediction accuracy ($r$, blue to red, **(a, d, g, j, m, p)**), logarithmic ridge hyperparameter ($\log_{10} \lambda$, gray to white, **(b, e, h, k, n, q)**), transfer function weights that are summed over delays ($b$, blue to green, **(c, f, i, l, o, r)**) are shown along the columns. The analysis was done in the 3-D space, but transverse slices (Montreal Neurological Institute [MNI]-coordinate Z = 8 mm) are chosen to display anatomical structures implicated in a meta-analysis on music-evoked emotions (Koelsch, 2020) such as the Heschl's gyrus, planum temporale, and the inferior frontal cortex. The 3-D volumes can be viewed with the NeuroVault web viewer (https://identifiers.org/neurovault.collection:19626).

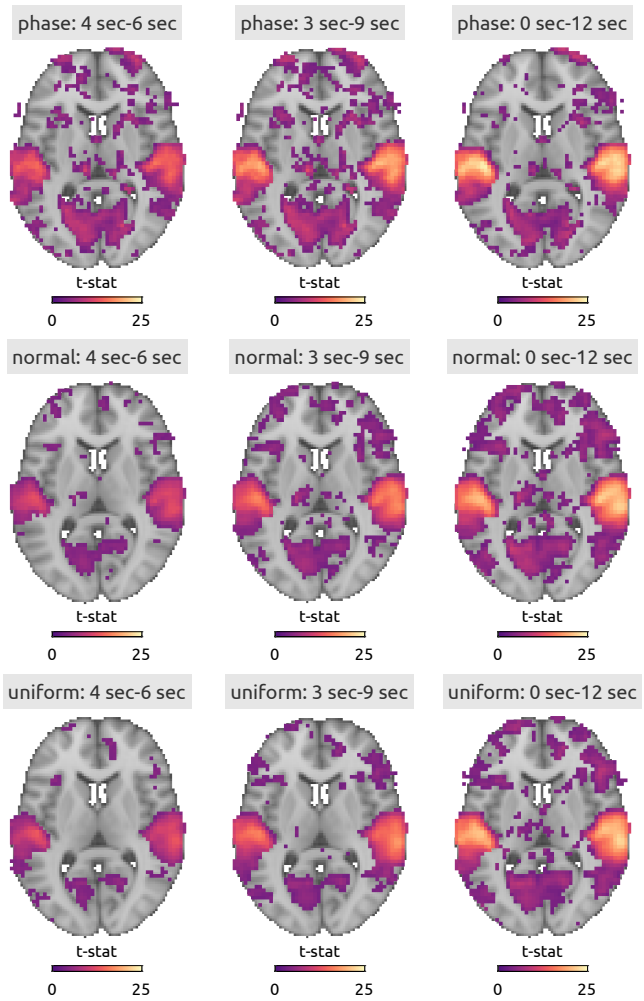


Figure 7: $t$-statistic maps on the transverse slices comparing the null prediction accuracies between two CV schemes (e.g., Figure 6**g** vs. Figure 4**p**) to test the SDL effects in the fMRI data with the phase-randomised envelope (top row), the normal noise (middle row), and the uniform noise (bottom row). Voxels were thresholded by statistical significance after FDR adjustment (one-sided $P_{FDR} < 0.01$). The background anatomical image is the MNI template included in FSL (`MNI152_T1_2mm_brain.nii.gz`). The 3-D volumes can be viewed with the NeuroVault web viewer (https://identifiers.org/neurovault.collection:19626).

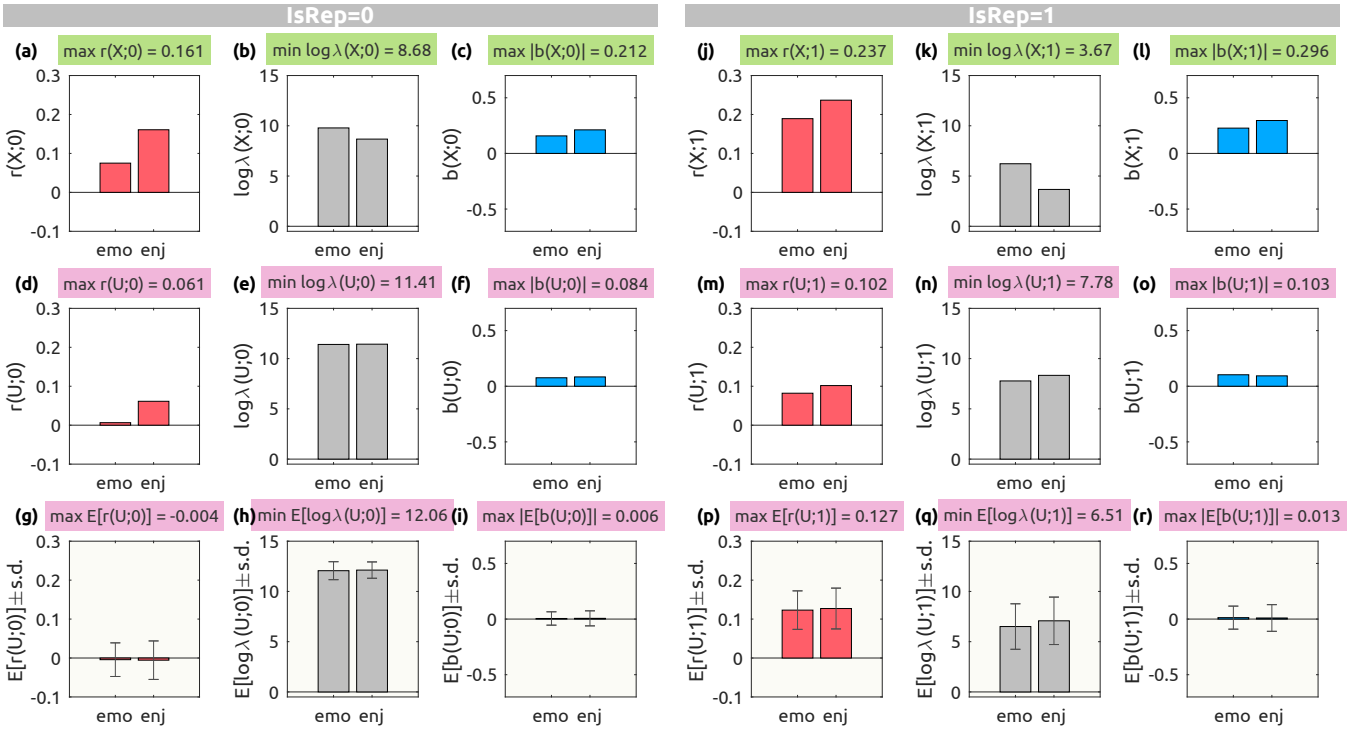


Figure 8: Behavioural linearised encoding analysis results with delays from 0 to 10 sec with an audio envelope (top row, **(a-c, j-l)**), a single case of the phase-randomised envelope (middle row, **(d-f, m-o)**), and an average of 100 phase-randomised envelopes (bottom row, pale yellow background, **(g-i, p-r)**). For each CV scheme ($\texttt{IsRep} = 0$, left panels, **(a-i)**; $\texttt{IsRep} = 1$, right panels, **(j-r)**), prediction accuracy ($r$, red bars, **(a, d, g, j, m, p)**), logarithmic ridge hyperparameter ($\log_{10} \lambda$, gray bars, **(b, e, h, k, n, q)**), transfer function weights that are summed over delays ($b$, blue bars, **(c, f, i, l, o, r)**) are shown along the columns. For the averaged metrics (bottom row), the standard deviations are shown as error bars. Emo.: emotionality, Enj.: enjoyment.

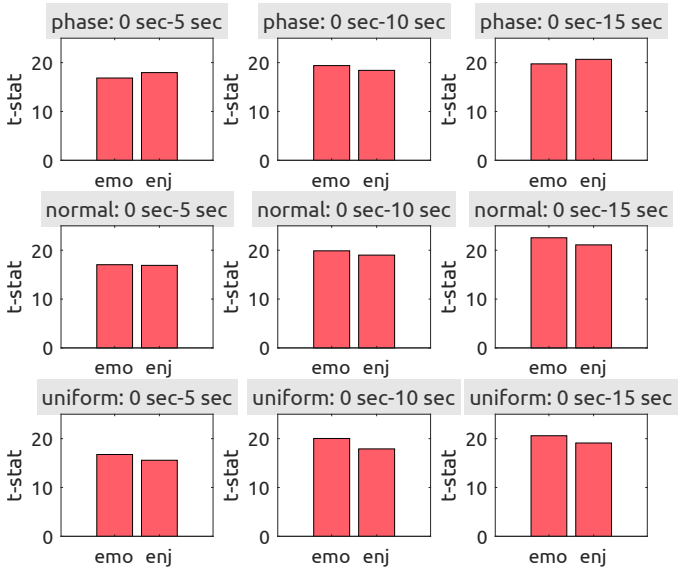


Figure 9: $t$-statistic bar plots comparing the null prediction accuracies between two CV schemes (e.g., Figure 8**g** vs. Figure 4**p**) to test the SDL effects in the behavioural data with the phase-randomised envelope (top row), the normal noise (middle row), and the uniform noise (bottom row). All effects were statistical significant after FDR adjustment (one-sided $P_{FDR} < 0.01$). Emo.: emotionality, Enj.: enjoyment.

prevent data leakage when testing the model's generalisability to unseen data. This paper elucidates how stimulus-driven leakage (SDL) can artificially inflate prediction accuracy and demonstrates realistic cases through simulations and real data analyses.

Firstly, it was mathematically shown that the expected prediction accuracy of the null feature could exceed zero when the null features are identically repeated across CV partitions (Equation 6). In particular, the similarity between training and validation sets disables regularisation leading to an inflation of the test prediction accuracy of the null features.

Secondly, simulations showed that the SDL artefact is more pronounced with a higher SNR, greater flexibility (i.e., more delay points or higher dimensional features), and more similar autocorrelation structures between the true and null features (Supplementary Simulation results).

Lastly, the SDL artefact was consistently observed across popular data modalities in cognitive neuroscience (i.e., EEG, fMRI, behavioural ratings). In particular, the inflated prediction accuracy exhibited highly plausible spatial patterns even when predicted by uniform noise as a null feature. These patterns are driven by time-locked neural responses to repeated stimuli (widely known as inter-trial/-subject synchrony), not by random noise in the data. Therefore, when combined with *informal reverse inference* (i.e., inferring a mental process from a brain activity pattern without accounting for how likely that pattern is

to occur with other mental processes), which is also a common logical fallacy in cognitive neuroscience (Poldrack, 2006), SDL can lead to completely incorrect conclusions (e.g., a conclusion such as *"The auditory cortex was encoding this uniform random noise that was never presented to the participant."* based on Supplementary Figure 25).

## 5.1 But my features are not just random noise!

For demonstrative purposes, I used random noise to show that the SDL effect can be observed with a feature that is not expected to be encoded in the real data. Certainly, no one would sincerely expect the auditory cortex to encode a random feature that cannot be extracted from the stimulus. In practice, researchers hypothesise that certain information extracted from the stimulus is encoded in the response based on some theoretical and/or empirical grounds. However, SDL can inflate the prediction accuracy of their hypothesised feature significantly regardless of whether the information is actually encoded in the response. This, in turn, can contribute to contamination of the literature and misguide future research.

Also, note that the SDL artefact was greater for the phase-randomised envelope than the normal or uniform noise (Figure 5, Figure 7, Figure 9). This is because the phase-randomised envelope preserves the autocorrelation structure of the stimulus, making it more similar to true features than random noise. Since the hypothesised feature is extracted from actual stimuli, it necessarily bears some resemblance to true features, regardless of the non-linear operations involved. Therefore, the risk of SDL is greater when the hypothesised feature is derived from the stimulus rather than from random noise.

Often nested regression models are compared to determine the unique predictive contribution of a feature of interest. For example, in our previous study (Leahy et al., 2021), the prediction accuracy of a model with an audio envelope ("reduced model") was subtracted from the prediction accuracy of a model with the envelope and musical beats ("full model") to test the encoding of musical beats. While the comparison of nested models is a standard practice in Ordinary Least Squares (OLS) regression, it can be complicated with regularised regression such as ridge. The major problem is over-regularisation of relevant features due to irrelevant features in the full model when a single penalty hyperparameter is used for all features (feature spaces). A multi-penalty model can address this issue better, e.g., Kim et al. (2024) and La Tour et al. (2022).

However, even for model comparison, SDL can still inflate the prediction accuracy of the full model because the extra features would increase the flexibility of the model. Once again, in well-designed cross-validation without SDL, nonpredicitve features would have been regularised (i.e., regularising the model's flexibility). However, in the presence of SDL, the regularisation is disabled, and the nonpredictive

features are not penalised, thus leading to an inflated prediction accuracy.

## 5.2 What if I use null models for statistical inference?

In this paper, the statistical inference on the SDL effect was done by permutation test (i.e., randomly permuting the CV scheme labels). In practice, null features (or surrogate features) can be used for statistical inference to test whether the features of interest predict significantly better than the null features (e.g., Kaneshiro et al., 2020). This is particularly useful for naturalistic neuroimaging where the features and responses have strong autocorrelation and multicollinearity structures that need to be preserved while creating a null distribution. Under the null hypothesis (i.e., the feature of interest is not encoded in the neural data), the empirical regularisation is already at a sufficient level for a null feature. Thus, no additional regularisation is performed for null features over tens of thousands of iterations. However, if the null hypothesis is not true in presence of SDL, the entire null prediction accuracies will be inflated, shifting the null distribution along the positive direction. That is, even if the feature of interest is truly encoded in the neural data, the resulting nonparametric $P$-value would be overly pessimistic, leading to Type-II errors.

## 5.3 Is SDL relevant to other analyses?

As mentioned earlier, SDL is not unique to linearised encoding analysis or naturalistic neuroimaging. Researchers using other predictive analyses in cognitive neuroscience may therefore wonder whether SDL is relevant to their methods as well. In this section, I briefly review several methods and examine whether they could in principle be susceptible to SDL when applied to datasets with repeatedly presented stimuli. Note that the discussion here is not intended to evaluate or criticise specific published studies.

### 5.3.1 Beta image encoding

Unlike the FIR model that estimates transfer function weights for each time point, the beta image encoding model is on top of the classical GLM that estimates 'beta' weights for each stimulus. This approach is popular in visual fMRI experiments where the gazing and visual attention of participants is difficult (or unnecessary) to resolve in time (Kay et al., 2008) or auditory fMRI experiments with short (1–2 seconds) stimuli (Moerel et al., 2018). Thus, instead of directly handling the autocorrelated BOLD time series, an average activation amplitude (i.e., beta weight) during a short trial is first estimated using

a GLM, either using a theoretical transfer function called the canonical hemodynamic response function (**henson1999**) or by fitting a regularised FIR model on a split dataset (Prince et al., 2022).

The beta image encoding model is a linear model on the the estimated beta weights. For example, when $M$ stimuli (e.g., natural sounds) were used to estimate responses to $F$ features (e.g., spectrotemporal modulation rates), the linear model is given as

$$\xi = \mathbf{F}\mathbf{g} + \mathbf{e}, \tag{7}$$

where $\xi \in \mathbb{R}^{M\times 1}$ is the estimated stimulus-wise response vector (i.e., beta values of a certain voxel for $M$ stimuli), $\mathbf{F} \in \mathbb{R}^{M\times F}$ is a matrix that describes $F$ features for $M$ presented stimuli, $\mathbf{g} \in \mathbb{R}^{F\times 1}$ is a feature-wise weight vector of the voxel, and $\mathbf{e} \in \mathbb{R}^{M\times 1}$ is unknown noise.

In the beta image encoding, instead of the repetitions of time-locked signals, the repetitions of stimulus-locked signals would constitute SDL. For example, if two sets of beta images with $M$ identical stimuli were partitioned into CV folds (e.g., even runs vs. odd runs where all $M$ stimuli were presented in each run; i.e., the most common CV design for MVPA), the expected null prediction accuracy would be also positive:

$$\mathbb{E}\left[\text{corr}\left(\hat{\xi}_{2,\mathbf{H}},\, \xi_2\right)\right] \overset{(5)}{\approx} \mathbf{s}_1^{\mathsf{T}}\mathbf{H}_1\left(\mathbf{H}_1^{\mathsf{T}}\mathbf{H}_1\right)^{-1}\mathbf{H}_1^{\mathsf{T}}\mathbf{s}_1 \geq 0, \tag{8}$$

where $\mathbf{s}_i = \mathbf{F}_i\mathbf{g} \in \mathbb{R}^{M\times 1}$ is the underlying signal pattern (a "response profile") for $M$ stimuli in the $i$-th CV partition ($i = 1$, training; $i = 2$, testing), $\mathbf{H}_i \in \mathbb{R}^{M\times F}$ is the null feature matrix.

Stimulus-ordered response vectors may look unlikely because of the standard practice of randomising the presentation order of stimuli across runs and participants. However, the beta image estimation process can reorganize the response profiles. This realigns beta values in a consistent order across all runs (e.g., sequentially from the first to the $M$-th stimuli). Therefore, the risk of SDL in the beta image encoding analysis remains a concern, unless leave-$N$-stimulus-out CV (i.e., partitioning the $\xi$ vector into CV sets) is used.

### 5.3.2 Stimulus reconstruction

Reconstruction of unseen stimuli (e.g., images or sounds) based on neural data demonstrates the remarkable potential of neuroimaging techniques for "mind-reading" (Han et al., 2019; Kay et al., 2008; Santoro et al., 2017). In practice, a set of linear models decodes features from neural data (e.g., beta images), followed by a reconstruction step where a simple classifier or a deep neural network (such as variational autoencoder) synthesizes the stimulus from the decoded features.

Since the decoding model is also a linear model like the encoding model, in principle, SDL can also occur. However, the goal of the analysis makes the requirement of "unseen stimuli" more explicitly. For example, if one fits a linear model with $N-1$ subjects' responses to Stimulus A, and then tries to recover, once again, Stimulus A from a response of the $i$-th subject, the problem of non-independence between training and test sets may be more visible. Having said that, it is still possible that latent similarity of stimuli is overlooked. Especially if there are too many stimuli to manually inspect, it is possible that some stimuli are indeed non-identical but highly similar (e.g., utterances by the same speaker; repetitions of musical phrases). Checking inter-stimulus correlation for all features prior to partitioning them into training and test sets will prevent such cases of hidden similarity.

### 5.3.3 Multivariate classification

Multivariate classification analysis has been widely used in the cognitive neuroscience community, commonly known as multivoxel (or multivariate) pattern analysis (MVPA; Kriegeskorte et al., 2006) or single-trial classification in electrophysiological data such as EEG, MEG, and ECoG (Müller-Gerking et al., 1999; Pistohl et al., 2012; Quandt et al., 2012), either on the whole set of response units (e.g., whole-brain classification; Ryali et al., 2010) or local neighbors (e.g., a "searchlight"; Kriegeskorte et al., 2006).

Classification can be seen as model-free as compared to encoding or decoding (i.e., stimulus reconstruction) linear models because classification does not require a definition of features. The trained weight vector only represents a separation of given training examples in the response space, regardless of the physical or semantic features of the stimuli. Moreover, in principle, the repetition of stimuli (or at least classes) across data points is in fact necessary for the classification. In order to train a classifier, not only linear but also non-linear ones, balanced training and testing examples of all classes are required (Hastie et al., 2009). In other words, it is impossible to train a classifier for an 'unseen class'. A classifier can only classify an 'unseen instance' of a known class. Therefore, SDL does not apply to classification analysis, although other forms of data leakage—such as global preprocessing or feature selection before split—remain a great concern (Kapoor & Narayanan, 2023).

### 5.3.4 Representational similarity analysis

Finally, let us consider a method that evaluates the second-order isomorphism across representational systems (Kriegeskorte et al., 2008), widely known as representational similarity analysis (RSA). This flexible method defines a 'model' distance matrix between stimuli either based on class labels (as in

classification analysis) or feature descriptors (as in encoding analysis). The original formulation of RSA does not involve cross-validation, as the RSA itself is neither a predictive model nor a classification method (Kriegeskorte et al., 2008). A later extension introduced a cross-validated, squared Mahalanobis distance estimator—known as "crossnobis"—to enhance both reliability and interpretability (Diedrichsen & Kriegeskorte, 2017). When estimating a crossnobis distance between two conditions in a leave-one-out cross-validation (LOOCV) scheme, it is assumed that the number of responses for both conditions remains balanced across all CV partitions. Thus, if identical stimuli are repeated across partitions, the crossnobis distance for the neural representation distance matrix (RDM) would be biased by the stimulus-specific (rather than condition-specific) activity. While this is not over-fitting to identical noise, it would be over-fitting to the idiosyncratic signal of a particular stimulus. However, even in this case, null descriptors would define a model RDM that is irrelevant to the neural RDM. Therefore, a false conclusion that "a brain region represents null information" would not be supported by the association between the model and neural RDMs. Thus, the risk of SDL in RSA appears minimal.

## 5.4 How to detect and prevent SDL

Since data leakage in predictive analysis is a non-trivial problem, valuable pedagogical resources have been developed. For example, Bernett et al., 2024 proposes guiding questions that help researchers identify and avoid data leakage. Here, I discuss ways to detect SDL and types of alternative analyses and designs that can prevent it.

### 5.4.1 Algorithmic detection

As shown earlier, SDL occurs when the identical stimulus is presented more than once across CV partitions. Thus, a simple way to detect the risk of SDL would be looking for the identical (or similar) data pairs in the dataset, before partitioning the data into training, validation, and test sets.

This was already illustrated in the toy example (see Supplementary Results: Simulation). When no stimulus was repeated across CV partitions the intertrial correlation (ITC) was on average zero across 200 random samplings. However, when the stimulus was repeated across CV partitions, the ITC was on average about 0.8. Because this correlation between responses or features can be cheaply computed prior to costly optimisation and modelling fitting, this can be a useful diagnostic tool to assess risk for SDL.

Checking ITC is also useful to detect latent similarity across stimuli as well. For example, two audio files

are named differently but contain similar music, the researcher would not know about the presence of SDL. The feature-ITC can be a useful tool to detect such latent similarity.

For users' convenience, an automatic validation test for a given CV design based on feature-ITC and response-ITC is implemented as a default option in the MATLAB package for Linearised Encoding Analysis (LEA; https://github.com/seunggookim/lea).

### 5.4.2 Alternative analyses

If a risk of SDL is detected in your planned analysis, what can be done next? One way to handle it is to find an alternative CV design that does not introduce SDL.

**Subject-wise modelling** As shown in Figure 1, if possible, an easy solution is to adopt subject-wise modelling instead of stimulus-wise modelling. Unless there exists strong similarity between subjects (e.g., identical twins, i.e., literal 'twinning') in different CV partitions, subject-wise modelling can avoid SDL.

**Averaging responses** In practice, many researchers are motivated to aggregate the limited data across multiple subjects. This is because acquiring high-quality neuroimaging data remains time-consuming and expensive, while its SNR—particularly for non-invasive in-vivo methods—remains poor. Moreover, acquiring extensive data from vulnerable populations (e.g., patients, young children, elderly people) poses not only financial and logistical but also ethical concerns. Therefore, instead of subject-wise modelling, aggregating data across multiple subjects (i.e., stimulus-wise modelling) may seem to be reasonable.

If the noise ceiling (i.e., the SNR) is deemed undesirable for subject-wise modelling (see Lage-Castellanos et al., 2019 for noise ceiling estimation methods), one can average responses across subjects for each stimulus before the encoding analysis. For example, if an identical set of stimuli is used for all participants, one can average participants' responses for each stimulus to create an 'average-subject'. Then, subject-level inference can be performed (e.g., using null [surrogate] features that preserve autocorrelation, such as linear shifting or phase randomisation).

### 5.4.3 Alternative designs

Moreover, if this risk is recognised at an early stage of the study (e.g., during study design or after a few pilot sessions), one can consider alterative designs that deter SDL.

**Hold-out validation** A straightforward but expensive (i.e., requiring more data) way to prevent SDL is to use *hold-out validation* instead of cross-validation. In hold-out validation, the test set is used only once to estimate the prediction performance. Thus, the test set is deliberately designed to contain different stimuli, even before data acquisition. Because the testing accuracy is proportional to the signal strength of the test set (see Supplementary Equation 11), it is desirable to design the test set to have high signal strength (e.g., by averaging multiple presentations; Han et al., 2019; Huth et al., 2016; Nishimoto et al., 2011).

**Single-use stimulus** Another design to prevent SDL is to use a stimulus only once during the whole study. That is, a stimulus is presented to only one participant, only once, and never used again (i.e., a single-use stimulus). This design prevents any accidental stimulus repetition across CV partitions and thus avoids the risk of SDL in any CV designs. However, as mentioned above, the poor SNR of non-invasive neuroimaging data may need to be addressed. In that case, a stimulus may be presented multiple times to only one participant, but then the neuroimaging responses must be averaged across trials for each stimulus before any predictive analysis.

## 5.5 Limitations of the current paper

While the current paper provides a comprehensive analysis of SDL, some limitations should be acknowledged.

First, the paper focuses on a particular, relatively simple model—single-penalty ridge regression. That is, other types of regularisation (e.g., multi-penalty ridge, LASSO, elastic net) and non-linear modelling techniques (e.g., k-nearest neighbours regression, support vector regression, non-linear kernel ridge regression) were not explored. While the key mechanism of SDL (i.e., disabling regularisation due to the similarity in the underlying signal between the training and validation sets) is likely to be present in these other methods, determining the extent to which SDL influences them requires further analysis.

Second, no neural spike data or intracranial recordings—which are gradually becoming more accessible in human patients—were analysed in this paper. The SDL effect was only demonstrated in the context of continuous-valued data acquired from non-invasive methods (e.g., EEG, fMRI, and behavioural ratings). Sparse spike data or firing rate data may behave differently from continuous-valued data. However, given that the SDL effect is driven by similarity in the underlying signal, it is likely to occur when the raw data are processed so that stimulus-evoked, time-locked responses are strongly present (e.g., high-gamma power envelopes of local neuronal populations; Jacobs & Kahana, 2009; Ray et al., 2008).

## 5.6 Conclusion

The current paper shows that SDL is a critical but under-recognized risk in encoding analysis and stimulus reconstruction. SDL may lead to spurious conclusions suggesting that irrelevant information is encoded in neural signals. When carefully designed, the model-based approach remains a powerful information-based tool for naturalistic neuroimaging.

# Data and Code Availability

All analysis code, simulation data, and real data analysis results are available at Zenodo: https://doi.org/10.5281/zenodo.15100830. An executable demo of simulation is available at CodeOcean: https://codeocean.com/capsule/4591394/. The EEG raw dataset is available at Standford Digital Repository: https://purl.stanford.edu/sd922db3535. The fMRI and behavioural raw datasets are available at OpenNeuro: https://openneuro.org/datasets/ds003085. A MATLAB package for Linearised Encoding Analysis (LEA) on multimodal data is available at Zenodo release: https://doi.org/10.5281/zenodo.15107756 and GitHub: https://github.com/seunggookim/lea. A freely executable demo of LEA on real data is available at MATLAB Online: https://s.gwdg.de/7cOQmw. fMRI analysis results in 3-D NIfTI format can be viewed and downloaded at NeuroVault: https://identifiers.org/neurovault.collection:19626.

# Author Contributions

*Seung-Goo Kim:* Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing—Original Draft, Writing—Review & Editing, Visualization.

# Funding

Max Planck Society supported this work.

# Declaration of Competing Interests

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Acknowledgements

The author thanks Dr. Daniela Sammler, Dr. Vincent Shi-chen Chien, Mr. Seung-Cheol Baek, and three anonymous reviewers for their constructive feedback and expert insights on earlier versions of the manuscript.

# References

Bernett, J., Blumenthal, D. B., Grimm, D. G., Haselbeck, F., Joeres, R., Kalinina, O. V., & List, M. (2024). Guiding questions to avoid data leakage in biological machine learning applications. *Nature Methods*, *21*(8), 1444–1453.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906. https://doi.org/10.1121/1.1945807

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *Volume 10 - 2016*.

Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., & Lalor, E. C. (2021). Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research. *Frontiers in neuroscience*, *15*, 705621.

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS computational biology*, *13*(4), e1005508.

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, *33*(13), 5728–5735.

Dupré la Tour, T., Visconti di Oleggio Castello, M., & Gallant, J. L. (2025). The voxelwise encoding model framework: A tutorial introduction to fitting encoding models to fmri data. *Imaging Neuroscience*, *3*, imag_a_00575. https://doi.org/10.1162/imag_a_00575

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582. https://doi.org/10.1080/23273798.2018.1499946

Han, K., Wen, H., Shi, J., Lu, K.-H., Zhang, Y., Fu, D., & Liu, Z. (2019). Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, *198*, 125–136.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Jacobs, J., & Kahana, M. J. (2009). Neural representations of individual stimuli in humans revealed by gamma-band electrocorticographic activity. *Journal of neuroscience*, *29*(33), 10203–10214.

Kaneshiro, B., Nguyen, D. T., Norcia, A. M., Dmochowski, J. P., & Berger, J. (2020). Natural music evokes correlated eeg responses reflecting temporal structure and beat. *NeuroImage*, *214*, 116559. https://doi.org/10.1016/j.neuroimage.2020.116559

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9).

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data*, *6*(4). https://doi.org/10.1145/2382577.2382579

Kay, K. N. (2018). Principles for models of neural information processing. *NeuroImage*, *180*, 101–109.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Kim, S.-G. (2022). On the encoding of natural music in computational models and human brains. *Frontiers in Neuroscience*, *16*. https://doi.org/10.3389/fnins.2022.928841

Kim, S.-G., De Martino, F., & Overath, T. (2024). Linguistic modulation of the neural encoding of phonemes. *Cerebral Cortex*, *34*(4), bhae155.

Koelsch, S. (2020). A coordinate-based meta-analysis of music-evoked emotions. *NeuroImage*, *223*, 117350.

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

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

La Tour, T. D., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, *264*, 119728.

Lage-Castellanos, A., Valente, G., Formisano, E., & De Martino, F. (2019). Methods for computing the maximum performance of computational models of fmri responses. *PLOS Computational Biology*, *15*(3), 1–25.

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology*, *102*(1), 349–359.

Leahy, J., Kim, S.-G., Wan, J., & Overath, T. (2021). An analytical framework of tonal and rhythmic hierarchy in natural music using the multivariate temporal response function. *Frontiers in Neuroscience*, *15*, 665767.

Moerel, M., De Martino, F., Kemper, V. G., Schmitter, S., Vu, A. T., Uğurbil, K., Formisano, E., & Yacoub, E. (2018). Sensitivity and specificity considerations for fmri encoding, decoding, and mapping of auditory cortex at ultra-high field. *NeuroImage*, *164*, 18–31.

Müller-Gerking, J., Pfurtscheller, G., & Flyvbjerg, H. (1999). Designing optimal spatial filters for single-trial eeg classification in a movement task. *Clinical Neurophysiology*, *110*(5), 787–798.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fmri. *Neuroimage*, *56*(2), 400–410.

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254. https://doi.org/https://doi.org/10.1016/j.neuroimage.2020.117254

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

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

Pistohl, T., Schulze-Bonhage, A., Aertsen, A., Mehring, C., & Ball, T. (2012). Decoding natural grasp types from human ecog. *NeuroImage*, *59*(1), 248–260.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in cognitive sciences*, *10*(2), 59–63.

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fmri response estimates using glmsingle. *Elife*, *11*, e77599.

Quandt, F., Reichert, C., Hinrichs, H., Heinze, H., Knight, R., & Rieger, J. (2012). Single trial discrimination of individual finger movements on one hand: A combined meg and eeg study. *NeuroImage*, *59*(4), 3316–3324.

Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J., & Hsiao, S. S. (2008). Neural correlates of high-gamma oscillations (60–200 hz) in macaque local field potentials and their potential implications in electrocorticography. *Journal of Neuroscience*, *28*(45), 11526–11536.

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, *15*(1), 1829.

Ryali, S., Supekar, K., Abrams, D. A., & Menon, V. (2010). Sparse logistic regression for whole-brain classification of fmri data. *NeuroImage*, *51*(2), 752–764.

Sachs, M. E., Habibi, A., Damasio, A., & Kaplan, J. T. (2020). Dynamic intersubject neural synchronization reflects affective responses to sad music. *NeuroImage*, *218*, 116512. https://doi.org/10.1016/j.neuroimage.2019.116512

Santoro, R., Moerel, M., Martino, F. D., Valente, G., Ugurbil, K., Yacoub, E., & Formisano, E. (2017). Reconstructing the spectrotemporal modulations of real-life sounds from fmri response patterns. *Proceedings of the National Academy of Sciences*, *114*(18), 4799–4804.

Seitz-Holland, J., Haas, S. S., Penzel, N., Reichenberg, A., & Pasternak, O. (2024). Brainage, brain health, and mental disorders: A systematic review. *Neuroscience & Biobehavioral Reviews*, *159*, 105581.

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/https://doi.org/10.1016/j.tics.2019.05.004

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of Neuroscience*, *20*(6), 2315–2331. https://www.jneurosci.org/content/20/6/2315

Verstynen, T., & Kording, K. P. (2023). Overfitting to 'predict'suicidal ideation. *Nature Human Behaviour*, *7*(5), 680–681.

644 Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory
645 neurons by system identification. *Annu. Rev. Neurosci.*, *29*(1), 477–505. https://doi.org/10.
646 1146/annurev.neuro.29.051605.113024