Probabilistic Forecasting and Information-Theoretic Analysis of Multivariate fMRI Dynamics

Arda Bayer * , Zhiyao Zhang , Ahmet Emre Ipek , Rose Khavari , Behnaam Aazhang

Posted Date: 27 May 2026

doi: 10.20944/preprints202605.1857.v1

Keywords: functionalmagneticresonanceimaging; BOLDsignal; probabilisticforecasting; information theory; entropy; directed information; brain dynamics; stochastic processes; transformer models; recurrent neural networks

 doi:10.20944/preprints202605.1857.v1

Article

Probabilistic Forecasting and Information-Theoretic Analysis of Multivariate fMRI Dynamics

Arda Bayer [1,*], Zhiyao Zhang [1], Ahmet Emre Ipek [2], Rose Khavari [3] and Behnaam Aazhang [1]

[1] Department of Electrical & Computer Engineering, Rice University, TX 77005, USA
[2] Department of Electrical & Electronics Engineering, Özyeğin University, 34794 Istanbul, Türkiye
[3] Department of Urology, Houston Methodist, TX 77030, USA
* Correspondence: arda.bayer@rice.edu

Abstract

Functional magnetic resonance imaging (fMRI) signals exhibit complex temporal structure arising from multivariate neural dynamics, physiological variability, and measurement uncertainty. In this work, we formulate region-of-interest-level fMRI analysis as a probabilistic multi-step forecasting problem and investigate the predictability of blood-oxygen-level-dependent (BOLD) activity from an information-theoretic perspective. Using the Natural Scenes Dataset, we model multiregional BOLD activity as a stochastic process with finite memory and train multiple forecasting architectures, including linear regression, exponential smoothing, recurrent neural networks, and transformer-based models, to predict future BOLD samples from preceding temporal observations. Forecasting performance is analyzed together with entropy-based quantities, including marginal entropy, conditional entropy, and normalized predictive information measures estimated directly from model-derived predictive distributions without imposing restrictive Gaussian assumptions on the underlying BOLD dynamics. The best-performing model achieved significant improvement over a naive persistence baseline ($p = 0.001$) while yielding a high predictive information fraction ($\eta = 75.49\%$). Post hoc directed information analysis revealed that short-horizon prediction was dominated primarily by autoregressive, within-ROI, temporal structure. Overall, the proposed framework demonstrates how probabilistic forecasting and information-theoretic analysis can be integrated to characterize the predictability, uncertainty structure, and directional organization of large-scale fMRI dynamics, and may support future downstream neuroengineering and neural-state inference applications.

Keywords: functional magnetic resonance imaging; BOLD signal; probabilistic forecasting; information theory; entropy; directed information; brain dynamics; stochastic processes; transformer models; recurrent neural networks

1. Introduction

Forecasting is fundamentally linked to the amount of information available about a dynamical system. A system that can be predicted accurately over future time horizons necessarily contains structured temporal dependencies that reduce uncertainty about future states. From an information-theoretic perspective, forecastability therefore provides an empirical measure of the complexity, organization, and predictability of the underlying process. Beyond characterization alone, accurate forecasting is also closely related to downstream tasks such as monitoring, control, and decision making, where future system behavior must be estimated from incomplete observations.

In neuroscience, the ability to model and predict brain activity has become increasingly important with the rapid growth of brain-computer interfaces, neural decoding systems, and data-driven neurotechnology [1]. Recent advances in machine learning have demonstrated that complex cognitive and perceptual information can be inferred from neuroimaging signals, particularly functional magnetic resonance imaging (fMRI). For example, multiple studies have shown that visual stimuli[2,3], semantic

 doi:10.20944/preprints202605.1857.v1

representations, and even continuous language content can be reconstructed from fMRI recordings using large-scale predictive models and deep learning frameworks [4]. These include image reconstruction approaches based on latent diffusion and self-supervised representations, as well as semantic decoding systems capable of recovering continuous language representations from non-invasive brain recordings [3,5]. Collectively, these works demonstrate that fMRI blood-oxygen-level-dependent (BOLD) activity contains substantial latent structure that can support complex predictive and decoding tasks.

At the same time, comparatively less attention has been devoted to directly characterizing the intrinsic forecastability of fMRI dynamics themselves from an information-theoretic perspective. Most existing neuroimaging forecasting approaches are application-driven and focus primarily on reconstruction or decoding accuracy [2,5]. In contrast, forecasting performance can also be interpreted as a probe of the statistical organization of the BOLD signal and the extent to which future neural activity is constrained by prior temporal context.

In this work, we formulate multiregional fMRI forecasting as a probabilistic time-series prediction problem grounded in information theory. We model the BOLD signal as a multivariate stochastic process with finite temporal memory. Specifically, the process is represented using a memory parameter M and a prediction horizon H, where the task is to forecast the next H future samples using the preceding M observations across a set of predefined regions of interest (ROIs). Under this formulation, forecasting quality becomes directly related to the reduction in uncertainty achievable from past observations.

Using this framework, we investigate both the predictability and the information structure of ROI-level BOLD dynamics obtained from the Natural Scenes Dataset (NSD), a large-scale 7T fMRI dataset acquired during continuous natural scene viewing. We evaluate multiple forecasting architectures spanning classical statistical methods and modern machine learning approaches, including linear regression, exponential smoothing, recurrent neural networks, and transformer-based sequence models. Rather than focusing exclusively on predictive accuracy, we analyze forecasting performance through entropy-based quantities derived from the learned probabilistic models, including marginal entropy, conditional entropy, and normalized predictive information measures. In this setting, the forecasting models can also be interpreted as data-driven forward models of short-timescale brain dynamics, providing compact probabilistic representations of how future neural activity evolves from preceding observations. Such forward-model formulations are broadly relevant to downstream neuroengineering and neural-state inference problems, where future brain activity must be estimated from incomplete or noisy observations.

In addition to evaluating forecasting performance, we further investigated how predictive information was distributed across cortical regions using directed information (DI). Whereas forecasting accuracy alone quantifies how well future BOLD activity can be predicted, DI provides a complementary view into which ROI histories contribute most strongly to those predictions and how predictive structure propagates across the network. In this sense, the forecasting models can be interpreted not only as predictive tools but also as data-driven probes of directional statistical organization in multivariate brain dynamics. Directed information extends mutual information to temporally ordered stochastic processes and quantifies the extent to which the past activity of one process contributes to predicting the future activity of another process beyond self-history effects [6]. Related information-theoretic frameworks for dependence and information flow analysis have been studied extensively in information theory and network inference, including work associated with directed information and causal dependence measures in stochastic dynamical systems [7–9]. In the present study, DI was estimated post hoc from the trained probabilistic forecasting models, enabling directional analysis of predictive dependencies directly from the learned forecasting distributions.

The forecasting perspective adopted here is also motivated by broader developments in predictive modeling and sequential inference. In large-scale forecasting benchmarks such as the M5 forecasting competition, forecasting accuracy has been shown to depend strongly on the interaction between model

 doi:10.20944/preprints202605.1857.v1

structure, uncertainty estimation, and temporal dependencies in the underlying data[10]. Similar principles arise in neural systems, where the ability to predict future states may provide insight into the complexity and organization of brain dynamics [11,12]. By combining probabilistic forecasting with entropy and directed information analysis, the present work aims to provide a unified framework for studying the predictability, uncertainty structure, and directional organization of multivariate fMRI activity, while also motivating forecasting-based forward models for downstream neuroengineering and neural-state inference applications.

2. Materials and Methods

This section presents the dataset, preprocessing procedures, forecasting models, and information-theoretic analysis framework used in this study. Forecastability and directed dependencies were analyzed from an information-theoretic perspective, building on broader frameworks for inference and dependence analysis in complex stochastic systems [6,13].

2.1. Dataset and Region-of-Interest Definition

Functional magnetic resonance imaging (fMRI) data were obtained from the NSD [14], a large-scale 7T fMRI dataset acquired during continuous visual stimulation with natural scene images. NSD contains one of the largest amounts of densely sampled single-subject fMRI data currently publicly available, enabling high-resolution analysis of long-timescale cortical dynamics. In this work, we used the publicly released preprocessed 1.8 mm isotropic BOLD timeseries provided as part of the NSD preprocessing pipeline. Data from eight subjects were included. The NSD dataset is organized hierarchically into imaging sessions and individual functional runs, where each session contains multiple continuous BOLD acquisitions corresponding to separate stimulus presentation blocks. The present study focused on the visual cortex due to its strong stimulus-driven responses, well-characterized hierarchical organization, and suitability for studying multiscale predictive structure in brain dynamics.

A total of 23 visual cortical ROIs were defined in Montreal Neurological Institute (MNI) space using spherical masks centered on canonical visual-system coordinates. ROIs were defined as spherical regions centered on approximate MNI coordinates corresponding to established visual cortical areas identified in probabilistic retinotopic atlases and prior functional localization studies [15–18]. The primary visual cortex V1 was represented by a single midline ROI, whereas all remaining regions were represented bilaterally, resulting in 23 ROIs in total. Early visual cortex regions, V1, V2, V3, were assigned a radius of 5 mm, while higher-order visual areas used a radius of 6 mm as described in Table 1.

The ROI set spans multiple stages of the human visual system, including early retinotopic cortex (V1–V3), intermediate dorsal and ventral retinotopic areas (V3A, V3B, and hV4), lateral occipital retinotopic maps (LO1 and LO2), ventral occipital maps (VO1 and VO2), and category-selective ventral temporal regions, including the parahippocampal place area (PPA) and fusiform face area (FFA). The ROI coordinates were selected based on established visual neuroanatomical landmarks and prior retinotopic mapping literature[15–18].

To obtain subject-specific ROI timeseries, each ROI atlas defined in MNI space was nonlinearly warped into each subject's anatomical and functional space using ANTs-based registration [19]. Specifically, MNI-to-subject anatomical registration used symmetric normalization, followed by affine alignment between the subject anatomical and functional scan coordinate spaces. Mean BOLD activity within each ROI was then extracted for every timepoint of every run.

 doi:10.20944/preprints202605.1857.v1

Table 1. Regions of interest and their description.

Index (-/+)	ROI	MNI (x, y, z)	Radius
1	V1	$(0, -90, 0)$	5
2/3	V2	$(\pm 10, -85, 5)$	5
4/5	V3	$(\pm 15, -80, 10)$	5
6/7	hV4	$(\pm 25, -75, -10)$	6
8/9	V3A	$(\pm 20, -85, 25)$	6
10/11	V3B	$(\pm 25, -80, 30)$	6
12/13	LO1	$(\pm 35, -75, -5)$	6
14/15	LO2	$(\pm 40, -75, -5)$	6
16/17	VO1	$(\pm 25, -70, -15)$	6
18/19	VO2	$(\pm 30, -65, -15)$	6
20/21	PPA	$(\pm 28, -45, -12)$	6
22/23	FFA	$(\pm 40, -55, -15)$	6

2.2. Forecasting Problem Formulation

Let $X_t \in \mathbb{R}^N$ denote the multivariate BOLD signal across N ROIs at discrete time index t. Throughout this work, uppercase symbols such as X_t denote random variables or stochastic processes, whereas lowercase symbols such as x_t denote observed realizations of those variables. In addition, Y is used to denote the target random variable corresponding to future values of the underlying BOLD process X_t. Given a memory length $M \in \mathbb{Z}^+$, a positive integer, and forecasting horizon $H \in \mathbb{Z}^+$, the objective is to learn a forecasting function $f : \mathbb{R}^{M \times N} \to \mathbb{R}^{H \times N}$ that maps a window of past BOLD observations $(X_{t-M+1}, ..., X_t)$ to a prediction of future activity $(X_{t+1}, ..., X_{t+H})$. The forecasting problem was therefore formulated as supervised multivariate sequence prediction. Forecasting performance was interpreted as an empirical probe of the predictability and information structure of fMRI dynamics for the selected ROI.

Model performance was evaluated using multiple complementary metrics. Root mean squared error (RMSE) was used as the primary signal reconstruction metric. We additionally evaluated a scaled forecasting error Root Mean Squared Scaled Error (RMSSE) metric introduced in the M5 forecasting competition [10], where prediction error is normalized relative to a naive one-step forecasting baseline.

The normalized predictive information fraction was defined as

$$\eta = 1 - \frac{H(Y|X)}{H(Y)} = \frac{I(X;Y)}{H(Y)}, \tag{1}$$

which corresponds to an asymmetric normalized mutual information measure related to the uncertainty coefficient in information theory [20,21]. Here, $Y = (X_{t+1}, \ldots, X_{t+H})$ is the random variable corresponding to the future BOLD signal over the forecasting horizon H, and $X = (X_{t-M+1}, ..., X_t)$ is the random variable associated with the past observed signal window. For notational simplicity, entropy and mutual information quantities are written without explicit time dependence, implicitly assuming time-invariant joint statistics over the sampled forecasting windows. Here, $H(\cdot)$ denotes Shannon entropy and $H(\cdot|\cdot)$ the conditional Shannon entropy and $I(X;Y)$ denotes the mutual information between past and future BOLD activity [20]. By definition, $\eta \in [0, 1]$, where larger values indicate greater predictability of future signal dynamics from past observations. Equivalently, η can be interpreted as the relative reduction in expected coding length of the target signal Y when the past signal X is known compared to when it is unknown.

The marginal entropy and conditional entropy were estimated as

$$H(Y) \approx \mathbb{E}[-\log p_\theta(Y)], \qquad H(Y|X) \approx \mathbb{E}[-\log p_\theta(Y|X)], \tag{2}$$

 doi:10.20944/preprints202605.1857.v1

where $p_\theta(Y)$ and $p_\theta(Y|X)$ denote the discretized emprical marginal and discretized model predicted conditional probability distributions, respectively. The marginal distribution $p_\theta(Y)$ was estimated using a histogram-based discrete density estimator with 100 bins fit on the training portion of the dataset. To estimate the conditional distribution $p_\theta(Y|X)$, predictive residuals obtained from the trained forecasting models were discretized using the same histogram binning procedure. Let $\hat{Y} = f_\theta(X)$ denote the model prediction and define the residual variable $R = Y - \hat{Y}$. Empirical residual histograms were estimated independently for each ROI and forecasting horizon from calibration residuals obtained on held-out validation data. The conditional predictive distribution was then approximated as

$$p_\theta(Y|X) \approx p_R(Y - f_\theta(X)), \tag{3}$$

where $p_R(\cdot)$ denotes the discretized empirical residual distribution. This histogram-based construction enabled nonparametric approximation of conditional predictive likelihoods without imposing Gaussian assumptions on the forecasting residuals. Using the model-derived conditional distributions enabled estimation of η directly from the predictive structure learned by the forecasting models, thereby characterizing the information captured by the fitted dynamical representations.

2.3. Data Parsing and Sliding-Window Construction

Each fMRI run was independently normalized using run-level z-score normalization applied separately to each ROI timeseries:

$$\tilde{x}_t^i = \frac{x_t^i - \mu_i}{\sigma_i + \epsilon}, \tag{4}$$

where x_t^i is the BOLD signal for ROI i at discrete time t, μ_i and σ_i denote the within-run[14] mean and σ_i standard deviation of ROI i. This normalization avoids distribution leakage across subjects during leave-one-subject-out evaluation.

The normalized timeseries were transformed into supervised learning samples using an overlapping sliding-window procedure. For each run, input windows of length M and target windows of length H were extracted using an overlapping sliding-window procedure with one-sample increments. Specifically,

$$\mathbf{x}_t = [\tilde{x}_t, ..., \tilde{x}_{t+M-1}] \quad \text{and} \quad \mathbf{y}_t = [\tilde{x}_{t+M}, ..., \tilde{x}_{t+M+H-1}] \tag{5}$$

where $\tilde{x}_t$ is the normalized BOLD signal for N ROIs. Aggregating the windowed samples extracted across all NSD runs and sessions yielded the supervised forecasting dataset consisting of the input tensor $\mathbf{x}$ and target tensor $\mathbf{y}$.

This procedure converts each continuous multivariate BOLD sequence into a large collection of partially overlapping forecasting examples. Subject identities were preserved during splitting to prevent information leakage between training and testing partitions.

2.4. Forecasting Models

Four forecasting approaches spanning classical statistical models and modern deep learning architectures were evaluated.

2.4.1. Linear Regression

A ridge-regularized multivariate linear regression model [22] was used as one of the primary forecasting baselines. For each forecasting sample, the input window of shape $M \times N$ was flattened into a single feature vector of dimension MN, allowing the model to jointly learn temporal and cross-ROI relationships from the past BOLD activity. The regression model then learned a direct mapping from the flattened input history to the future forecasting target over the prediction horizon.

The linear forecasting model was parameterized by a weight tensor $W \in \mathbb{R}^{MN \times HN}$ defining a linear mapping f_W from the input history tensor $\mathbf{x}_t \in \mathbb{R}^{M \times N}$ to the future target tensor $\mathbf{y}_t \in \mathbb{R}^{H \times N}$.

 doi:10.20944/preprints202605.1857.v1

Given a collection of n supervised forecasting samples $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^{n}$, ridge regularization was used to estimate the forecasting weights:

$$\hat{W} = \arg\min_{W} \sum_{t=1}^{n} ||\mathbf{y}_t - f_W(\mathbf{x}_t)||_F^2 + \alpha ||W||_F^2, \tag{6}$$

where $||\cdot||_F$ denotes the Frobenius norm and $\alpha = 1.0$ controls the regularization strength. Ridge regularization was chosen to stabilize estimation in the high dimensional forecasting setting, particularly since the number of temporal forecasting features grows proportionally with both the memory length M and the number of ROIs N.

Although substantially simpler than deep neural forecasting architectures, the linear regression model can still capture a considerable portion of the temporal structure present in the BOLD signals. Comparisons between the forecasting performance of the linear model and higher-capacity nonlinear models provide insight into the underlying signal complexity and noise characteristics of the data. In particular, similar performance between linear and nonlinear models may indicate that the observed dynamics are dominated by approximately linear dependencies or by variability arising from unobserved sources that cannot be effectively modeled even by overparameterized nonlinear architectures.

2.4.2. Exponential Smoothing

Exponential smoothing is a classical time-series forecasting approach that recursively estimates future observations using weighted averages of past measurements, assigning progressively greater weight to more recent samples. Compared to the neural network architectures evaluated in this work, exponential smoothing provides a substantially simpler statistical forecasting baseline primarily intended to capture short-timescale autoregressive structure in ROI activity.

Exponential smoothing model, together with appropriate trend and seasonality extensions, served as strong baselines in the M5 forecasting competition [10], where the relatively simple statistical method were shown to outperform many more elaborate machine learning approaches [10]. Since the fMRI BOLD signals analyzed in this study did not exhibit a consistent global trend or periodic seasonal structure, no explicit trend or seasonal components were included in the model. Consequently, the implemented formulation corresponds to the *Simple Exponential Smoothing* (SES) model [23], in which future predictions are generated solely from exponentially weighted averages of past observations.

The simple exponential smoothing forecastor can be described as [23]

$$\hat{x}_t = \ell_{t-1} \tag{7}$$

$$\ell_t = \alpha x_t + (1-\alpha)\ell_{t-1}, \tag{8}$$

where x_t is the observed signal value corresponding to the random variable X_t, ℓ_t is the latent variable, $\alpha \in [0,1]$ is the smoothing parameter and $\hat{x}_t$ is the model prediction. Once a prediction is obtained for time instance t, it was recursively used as the observed signal to infer the remaining future values $\hat{x}_{t+h}$ over the horizon H for $h = 1, ..., H-1$. To ensure comparability with the other models, the latent variable was initialized such that $\ell_{t-m} = 0$ for $m > M$, while ℓ_0 was set to the mean BOLD signal value computed from the training data.

Unlike the other forecasting models evaluated in this work, exponential smoothing does not naturally support multivariate inputs in which multiple ROI time series are jointly used to predict future activity. Consequently, a separate exponential smoothing model was fit independently for each ROI using only its own preceding temporal observations. This yields a comparatively low-complexity forecasting framework in which each ROI model is parameterized primarily by a single smoothing coefficient, α, controlling the relative weighting of recent versus past observations.

 doi:10.20944/preprints202605.1857.v1

2.4.3. Long Short-Term Memory Network

To capture nonlinear temporal dependencies, we implemented a multi-layer Long Short-Term Memory (LSTM) network [24]. The model receives an input tensor of shape (M, N), where M denotes the temporal memory window and N the number of ROIs. The architecture consisted of three recurrent LSTM layers with a hidden-state dimension of 512 units and inter-layer dropout probability of 0.5 to reduce overfitting. The recurrent layers were configured using batch-first ordering, enabling direct processing of ROI time-series windows.

The hidden state of the final recurrent layer at the last observed time point was passed to a fully connected projection layer that maps the latent representation into the full forecasting horizon. Specifically, the final linear layer outputs a vector of dimension $H \times N$, which is reshaped into a prediction tensor of shape (H, N) corresponding to simultaneous multi-step forecasts across all ROIs. The training protocol details including the training loss and the optimizer pick are given in section 2.5.

2.4.4. Transformer

We further evaluated a transformer-based sequence model employing self-attention mechanisms for long-range temporal dependency modeling[25]. The transformer architecture operated on input windows of shape (M, N) and first projected each ROI vector into a latent embedding space of dimension $d_{\text{model}} = 64$ using a learned linear projection layer. Learned positional embeddings were added to preserve temporal ordering information within the sequence.

The embedded sequence was processed using an encoder-only transformer composed of two stacked transformer encoder layers with four attention heads per layer and dropout probability 0.1. The encoder outputs contextualized latent representations for all temporal positions. The latent representation corresponding to the final input time point was then passed through a linear output layer producing a vector of dimension $H \times N$, which was reshaped into the multi-step prediction tensor of shape (H, N). This formulation enables the model to capture distributed temporal interactions between ROIs through self-attention on the concatenated input space, $\mathbb{R}^{M \times N}$, while directly generating the full forecasting horizon in a single forward pass. The training protocol for this neural network is also provided in section 2.5.

2.5. Training and Evaluation Protocol

Model evaluation followed a two-stage subject-level generalization protocol. First, two subjects were reserved as a fully held-out test set and excluded entirely from model selection and cross-validation procedures. The remaining six subjects were then used for leave-one-subject-out cross-validation (LOSO-CV), where in each fold one subject served as the validation subject while the remaining subjects were used for training. This LOSO framework enabled assessment of cross-subject generalization while reducing the risk of subject-specific overfitting. For neural forecasting models, the best-performing model weights from the LOSO folds, determined using validation performance and early stopping, were subsequently evaluated on the previously unseen held-out test subjects to obtain the final test results.

For deep learning models, training employed the AdamW optimizer[26] with an initial learning rate of 5×10^{-4} and weight decay of 10^{-5}. Adaptive learning-rate scheduling was performed with multiplicative decay factor 0.5 [27], while early stopping based on validation loss was used to reduce overfitting. The training objective combined a Huber reconstruction loss for the predicted BOLD time series [28] with an additional temporal-difference regularization term that penalizes discrepancies in first-order temporal dynamics between predicted and observed signals:

$$\mathcal{L}(x_t, \hat{x}_t; \delta) = \mathcal{L}_{Huber}(x_t, \hat{x}_t; \delta) + \alpha \mathcal{L}_{\Delta}(x_t, \hat{x}_t; \delta) \tag{9}$$

$$\mathcal{L}_{\Delta}(x_t, \hat{x}_t; \delta) = \mathcal{L}_{Huber}(x_t - x_{t-1}, \hat{x}_t - \hat{x}_{t-1}; \delta), \tag{10}$$

where $\alpha = 0.3$ controls the contribution of the temporal-difference regularization term, $\delta = 0.5$ denotes the Huber threshold parameter, and $\mathcal{L}_{\Delta}$ penalizes discrepancies between predicted and observed

 doi:10.20944/preprints202605.1857.v1

temporal differences across adjacent forecast steps. For completeness, the Huber reconstruction loss was defined as

$$\mathcal{L}_{Huber}(x_t, \hat{x}_t; \delta) = \frac{1}{H} \sum_{h=0}^{H-1} \phi_\delta(x_{t+h} - \hat{x}_{t+h}) \tag{11}$$

$$\phi_\delta(z) = \begin{cases} \frac{1}{2}z^2, & |z| \leq \delta \\ \delta(|z| - \frac{1}{2}\delta), & |z| > \delta \end{cases} . \tag{12}$$

where H denotes the forecasting horizon.

We systematically varied the memory parameter M and forecasting horizon H to characterize the temporal predictability structure of the fMRI signals. Model selection was based on cross-validated forecasting performance across LOSO folds.

Forecasting results were additionally compared against a naive persistence baseline that repeats the last observed frame over the prediction horizon. Statistical significance was assessed using a permutation test with 1000 permutations, where training targets were randomly permuted to generate a null distribution of forecasting performance.

2.6. Directed Information Post Analysis

To investigate directed dependencies between ROIs, we performed a post hoc DI analysis using the trained forecasting models. DI provides a directional information-theoretic measure quantifying the extent to which the past activity of one source random variable contributes to predicting the future activity of a target random variable [6]. In contrast to forecasting accuracy alone, which quantifies overall predictability of future BOLD activity, DI provides a complementary view into which ROI histories contribute most strongly to those predictions and how predictive structure propagates across the network. Related information-theoretic frameworks for analyzing dependence and information flow in complex systems have also been studied extensively within the information theory literature [13].

For each source ROI signal X^i and target ROI signal X^j, we estimated pairwise marginal directed information by comparing predictive likelihoods obtained from the full model against a reduced-input model in which the source ROI history was removed. Specifically, the estimator takes the form[6,7,29]

$$I(X^i \to X^j) = \mathbb{E}[\log p(X^j_{t+1}|\mathcal{H}_t) - \log p(X^j_{t+1}|\mathcal{H}_t^{-i})], \tag{13}$$

where $p(A|B)$ denotes the conditional probability distribution of a random variable A given another random variable B, $\mathcal{H}_t$ denotes the complete multivariate history and $\mathcal{H}_t^{-i}$ denotes the history with ROI i removed. As in the entropy analysis, the expectation was estimated over pooled forecasting samples under an implicit assumption of time-invariant joint statistics across the sampled windows. The probabilistic predictive distributions were constructed post hoc using empirical residual histogram models. Let $\hat{X}^j_{t+1} = f_\theta(\mathcal{H}_t)$ denote the model prediction for ROI j at forecasting horizon $h = 1$, and define the residual variable

$$R^j_t = X^j_{t+1} - \hat{X}^j_{t+1}. \tag{14}$$

For each ROI and forecasting horizon, empirical residual distributions were estimated from calibration residuals using histogram-based discrete density estimation. The conditional predictive distribution was then approximated as

$$p(X^j_{t+1}|\mathcal{H}_t) \approx p_R\Big(X^j_{t+1} - f_\theta(\mathcal{H}_t)\Big), \tag{15}$$

where $p_R(\cdot)$ denotes the empirical residual probability mass function estimated from the calibration data. This construction enabled nonparametric sample-based approximation of conditional predictive likelihoods without imposing Gaussian assumptions on the forecasting residuals, thereby al-

 doi:10.20944/preprints202605.1857.v1

lowing directed information estimation from arbitrary forecasting architectures. The whole forecasting pipeline is summarized in Figure 1.

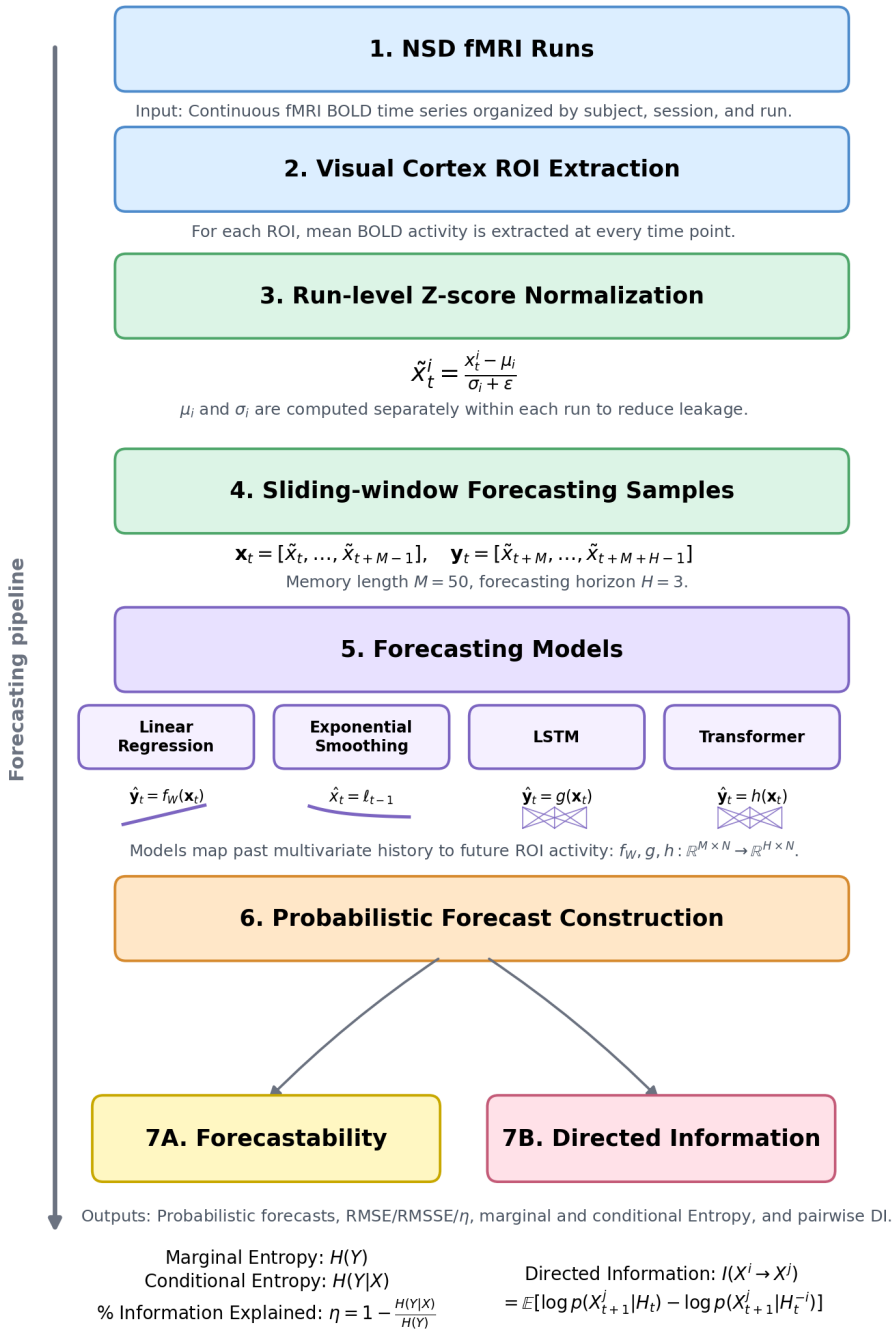


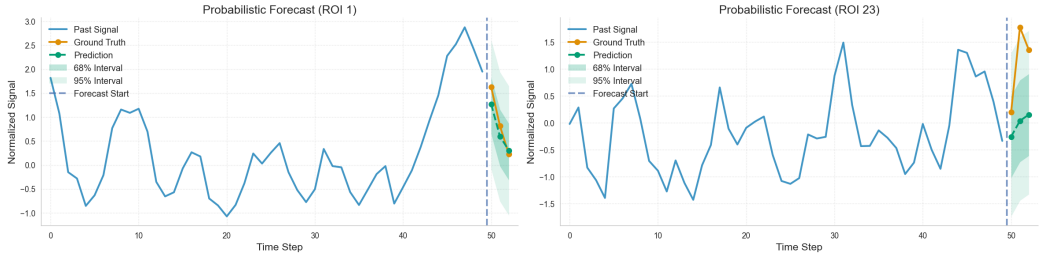
Figure 1. Overview of the proposed multivariate fMRI forecasting framework. Preprocessed BOLD timeseries extracted from predefined regions of interest (ROIs) were converted into sliding temporal windows and used to train forecasting models to predict future ROI activity from past observations. Forecasting performance and predictive distributions were subsequently used to compute error-based and information-theoretic measures, including conditional entropy and directed information.

3. Results

The multi-model forecasting framework was evaluated for multi-step prediction of ROI-level BOLD dynamics. Forecasting models were trained using leave-one-subject-out cross-validation with windowed BOLD sequences as input and future ROI activity as the prediction target.

 doi:10.20944/preprints202605.1857.v1

Figure 2 illustrates example three-step probabilistic forecasts generated by the transformer model for two representative ROIs. One example demonstrates relatively close agreement between the predicted and observed BOLD trajectories over the forecast horizon, whereas the second example highlights a case with larger prediction error. In both cases, the model provides a probabilistic estimate of future activity rather than a deterministic reconstruction of the signal. These examples emphasize that the forecasting framework captures statistical structure in the ROI dynamics while still exhibiting substantial uncertainty and imperfect predictive accuracy, consistent with the noisy and stochastic nature of fMRI BOLD measurements.



(**a**) (**b**)

Figure 2. Illustration of probabilistic forecasting performance for two representative ROIs. The historical ROI activity preceding the forecast horizon is shown together with the ground-truth future trajectory, model prediction, and uncertainty intervals estimated from forecasting residuals. **(a)** Example of a well-predicted ROI trajectory, where the model accurately captures the temporal evolution of the BOLD signal within the forecast horizon. **(b)** Example of a poorly predicted ROI trajectory, illustrating increased forecasting error and uncertainty. These examples highlight the heterogeneous forecastability of fMRI dynamics across ROIs and temporal instances.

Forecasting performance across the evaluated architectures is summarized in Table 2, including both cross-validation results and performance on the hold-out test subjects. The forecasting performances were comparable and the linear model achieved the best overall predictive performance among the validated approaches.

Table 2. Forecasting performance across models for cross-validation and hold-out test evaluation.

	Cross-Validation		**Hold-Out Test**	
Model	**RMSE**	**RMSSE**	**RMSE**	**RMSSE**
Naive Last Value	0.920	1.39	0.859	1.47
Linear Regression	**0.767**	**1.17**	**0.710**	**1.22**
Exponential Smoothing	0.891	1.35	0.846	1.45
LSTM	0.779	1.18	0.733	1.26
Transformer	0.770	**1.17**	0.721	1.23

Best results within each evaluation setting are shown in bold.

To quantify the information content captured by the forecasting framework, we computed entropy-based measures from the probabilistic forecasts. The transformer model achieved a mean marginal entropy of $H(Y) = 3.07$ nats and a mean conditional entropy of $H(Y|X) = 0.75$ nats across ROIs, yielding an average predictive information gain of $H(Y) - H(Y|X) = 2.32$ nats. The corresponding normalized predictive information fraction was $\eta = 75.49\%$ for the multivariate problem. Per-ROI entropy decomposition values are shown in Figure 3, where marginal entropy, conditional entropy, and normalized predictive information are visualized jointly for each ROI.

A permutation-based forecasting significance test using 1000 null-model realizations demonstrated that the proposed forecasting framework significantly outperformed the naive baseline (p=0.001).

 doi:10.20944/preprints202605.1857.v1

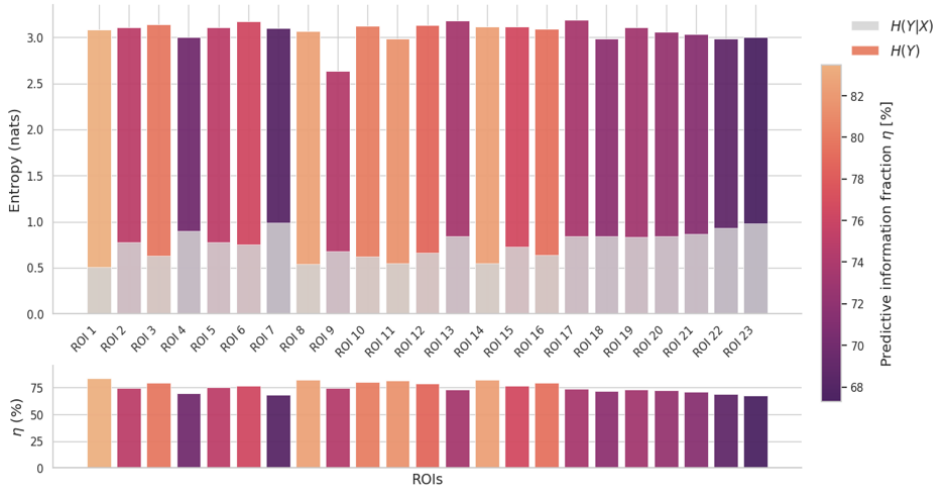


Figure 3. Forecastable information across ROIs. *Top Figure:* Barplot shows the information content of the target BOLD signal measured by empirical Shannon Entropy $H(Y)$. The conditional entropy $H(Y|X)$ of the target BOLD signal Y given past signal X is given in gray. The difference $H(Y) - H(Y|X)$ is the forecastable information and color codes the percent of forecastable information per ROI. *Bottom Figure:* Percent forecastable information $\eta = 1 - H(Y|X)/H(Y)$ per ROI.

3.1. Post Hoc Directed Information Analysis

Following model training, a post hoc DI analysis was performed using the probabilistic forecasting outputs. The DI computation compared predictive likelihoods obtained using the full ROI history against likelihoods obtained after selectively removing the history of individual source ROIs. This yielded a directed information matrix quantifying the contribution of each source ROI to predicting each target ROI.

The resulting DI matrix for the best-performing model is shown in Figure 4. For the selected memory length $M = 50$ and forecast horizon $H = 3$, the matrix exhibited a predominantly diagonal structure, indicating that the strongest predictive contributions arose from within-ROI temporal history rather than cross-ROI interactions.

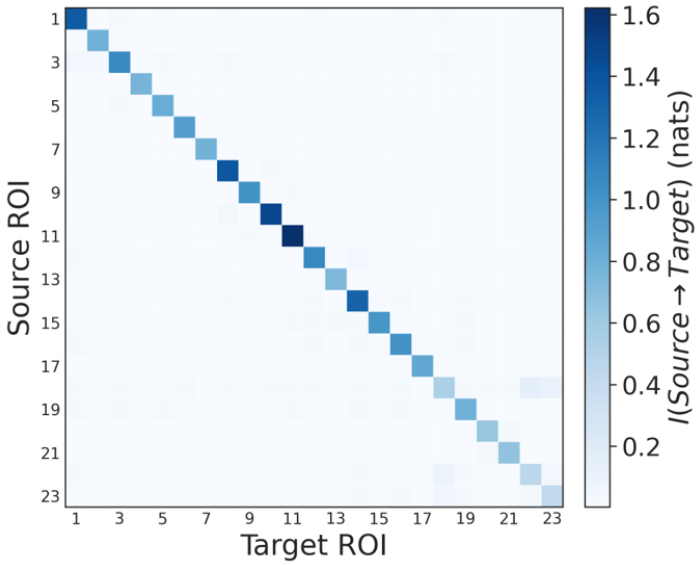


Figure 4. Directed information (DI) matrix estimated from the probabilistic forecasting model across the selected ROIs. Larger values indicate stronger directional predictive contributions from the source ROI (rows) to the target ROI (columns). The predominantly diagonal structure suggests that short-horizon BOLD forecasting is dominated primarily by autoregressive within-ROI temporal dependencies, whereas cross-ROI directed interactions remain comparatively weaker.

 doi:10.20944/preprints202605.1857.v1

4. Discussion

The present study demonstrates that multivariate fMRI BOLD dynamics contain substantial short-horizon predictive structure that can be captured using probabilistic forecasting models. Across all evaluated approaches, forecasting performance significantly exceeded that of a naive persistence baseline ($p = 0.001$), indicating that the models learned nontrivial temporal dependencies beyond simple signal persistence.

Entropy decomposition further showed that a substantial fraction of uncertainty in future BOLD activity could be reduced using preceding ROI history. The relatively low conditional entropy compared to the marginal entropy indicates that short-timescale fMRI dynamics are not purely stochastic, but instead exhibit measurable temporal organization. At the same time, the persistence of nonzero conditional entropy suggests that a considerable component of the signal remains unpredictable, potentially reflecting measurement noise, physiological variability, latent neural states, or unobserved network processes not represented within the selected ROI set.

The ROI-level entropy analysis additionally revealed heterogeneous predictive structure across the visual hierarchy. Marginal entropy values remained relatively consistent across ROIs, suggesting broadly similar overall signal variability throughout the visual cortex. In contrast, conditional entropy and predictive information fractions varied substantially between regions. Early and intermediate visual areas, including V1, V3A, V3B, and LO2, exhibited some of the highest predictability values, indicating strong local temporal continuity and relatively stable short-timescale dynamics. Conversely, higher-order ventral temporal regions, particularly the fusiform face area (FFA) and parahippocampal place area (PPA), demonstrated comparatively lower predictability and higher conditional entropy, suggesting increased stochastic variability or more complex latent dynamics. These findings are consistent with the hierarchical organization of the visual system, where early retinotopic cortex is more strongly constrained by local stimulus-driven temporal structure, whereas higher-order category-selective regions integrate more abstract and distributed representations [15,16].

Interestingly, forecasting performance was broadly comparable across both relatively simple and substantially more complex forecasting architectures. Linear regression, transformer, and LSTM models achieved similar validation behavior despite their differing representational capacities and modeling assumptions. This observation suggests that a large component of the short-horizon predictive structure in the examined BOLD signals may arise from relatively low-order temporal dependencies rather than highly nonlinear long-range interactions. From an information-theoretic perspective, the models also impose distinct assumptions regarding the underlying signal structure. Linear regression is naturally associated with approximately linear Gaussian dependencies, exponential smoothing emphasizes local temporal continuity, whereas neural architectures such as LSTMs and transformers can represent nonlinear and higher-dimensional temporal interactions. Nevertheless, all models remained constrained by the selected temporal memory window ($M = 50$) and forecasting horizon ($H = 3$), which define the effective temporal context available for prediction.

The post hoc directed information analysis further clarified the structure of the learned predictive dependencies. The predominantly diagonal DI matrices indicate that the future activity of a given ROI is primarily associated with its own recent temporal history rather than with the histories of other ROIs. This finding is consistent with the autoregressive nature of the best-performing forecasting models and suggests that short-horizon BOLD prediction in this setting is dominated largely by local temporal persistence. Notably, ROIs 10 and 11, corresponding to bilateral V3B, exhibited the largest diagonal DI values, indicating particularly strong self-predictive information within this intermediate dorsal visual region. This may reflect relatively stable local temporal structure or sustained stimulus-driven activity in V3B, which is known to participate in higher-order retinotopic processing and integration of visual motion and spatial information [15,16]. Interestingly, despite the well-established hierarchical organization of the visual cortex, the DI analysis did not reveal a strong feedforward or hierarchical off-diagonal interaction structure under the selected temporal memory and forecasting settings. Instead, the dominant predictive structure remained largely local and autoregressive. Nevertheless, smaller but

 doi:10.20944/preprints202605.1857.v1

nonzero off-diagonal DI values remained observable across several ROI pairs, including interactions between VO2 and FFA, suggesting weaker distributed statistical dependencies between higher-order ventral visual regions. These interactions may reflect broader network-level integration processes that are not fully captured by purely local autoregressive dynamics.

More broadly, the present framework illustrates how information-theoretic quantities can provide interpretable summaries of learned neural forecasting dynamics. Even for comparatively high-capacity models such as transformers, entropy decomposition and directed information analyses yielded interpretable measures of uncertainty reduction, predictability, and directional dependence. In this sense, the forecasting models function not only as predictive tools, but also as data-driven probes for characterizing the statistical organization of large-scale fMRI activity.

4.1. Limitations

This study has several limitations. Most notably, the forecasting horizon was limited to $H = 3$ future samples, since longer horizons resulted in substantially increased validation loss and unstable predictive performance. Consequently, the present findings primarily characterize short-timescale temporal dependencies in BOLD activity. Future work could investigate multiscale or hierarchical forecasting architectures capable of maintaining stable probabilistic predictions over longer temporal horizons, while also incorporating richer network-level or latent-state representations.

An additional limitation arises from the histogram-based discretization procedure used for estimating predictive probability distributions and entropy-based quantities. The underlying BOLD forecasting distributions are continuous-valued, whereas entropy and directed information estimation in the present study were performed using discretized empirical histogram approximations primarily to simplify likelihood estimation and numerical computation. As a consequence, the absolute entropy values reported in nats depend on the selected discretization resolution, specifically the $100-$bin histogram partition used throughout this work. In particular, finer histogram resolutions generally increase the estimated entropy values due to the increased partition cardinality. Accordingly, the reported entropy magnitudes should be interpreted relative to the discretization scheme employed rather than as absolute continuous entropy measures. Future work could instead employ continuous probabilistic density models and differential entropy estimators to obtain discretization-independent information-theoretic quantities.

5. Conclusions

In this work, we formulated multiregional fMRI BOLD forecasting as a probabilistic time-series prediction problem grounded in information theory. Using ROI-level activity from the Natural Scenes Dataset, we evaluated multiple forecasting architectures and quantified predictive structure through entropy-based measures and directed information analysis. The results demonstrated that short-horizon BOLD activity contains significant nontrivial temporal predictability beyond naive persistence baseline, while still retaining substantial stochastic variability.

Entropy decomposition revealed that a considerable fraction of future uncertainty could be reduced using recent ROI history, and post hoc directed information analysis indicated that short-timescale prediction was dominated primarily by local autoregressive structure. Importantly, the study also demonstrated that information-theoretic quantities can provide interpretable summaries of learned dynamics even for comparatively complex black-box forecasting models.

Overall, the presented framework connects probabilistic forecasting, information theory, and multivariate neuroimaging analysis within a unified setting. These results suggest that forecasting-based approaches may provide a useful tool for studying the predictability, complexity, and directional organization of large-scale brain dynamics.

Author Contributions: Conceptualization, A.B., B.A.; methodology, A.B., Z.Z. and A.I.; software, A.B., Z.Z., A.I; validation, A.B., Z.Z.; formal analysis, A.B., Z.Z., and A.I.; investigation, A.B., Z.Z. and A.I.; resources, A.B., B.A.; data curation, A.B.; writing—original draft preparation, A.B., Z.Z. and A.I.; writing—review and editing, A.B.,

 doi:10.20944/preprints202605.1857.v1

Z.Z., A.I., B.A.; visualization, A.B., Z.Z. and A.I.; supervision, A.B., B.A. and R.K.; project administration, A.B., B.A. and R.K.; funding acquisition, B.A. and R.K.. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Rice University Provost's TMC Collaborator Seed Fund.

Institutional Review Board Statement: This study used publicly available, de-identified data from the Natural Scenes Dataset (NSD) [14]. Ethical approval and informed consent procedures for data acquisition were conducted by the original NSD investigators as described in the original publication.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the original NSD study [14].

Data Availability Statement: The Natural Scenes Dataset (NSD) analyzed in this study is publicly available from the original NSD release [14]. The code used for forecasting, entropy estimation, and directed information analysis is publicly available at https://github.com/ab126/fmri_forecasting. An archived release of the code associated with this manuscript is available through Zenodo: https://doi.org/10.5281/zenodo.20341604

Acknowledgments: During the preparation of this study, the authors used GenAI [GPT 5.5, Codex, Claude Haiku 4.5, Gemini 3] for the initial implementation of the methods described herein and for the initial drafting of the manuscript. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BOLD	Blood-oxygen-level-dependent
CV	Cross-validation
DI	Directed Information
fMRI	Functional Magnetic Resonance Imaging
GenAI	Generative Artificial Intelligence
LOSO	Leave-one-subject-out
LSTM	Long Short-Term Memory
MNI	Montreal Neurological Institute
NSD	Natural Scenes Dataset
ROI	Region of Interest

References

1. Lebedev, M.A.; Nicolelis, M.A. Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation. *Physiological reviews* **2017**, *97*, 767–837.
2. Beliy, R.; Gaziv, G.; Hoogi, A.; Strappini, F.; Golan, T.; Irani, M. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems* **2019**, *32*.
3. Lin, S.; Sprague, T.; Singh, A.K. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems* **2022**, *35*, 29624–29636.
4. Tang, J.; LeBel, A.; Jain, S.; Huth, A.G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience* **2023**, *26*, 858–866.
5. Takagi, Y.; Nishimoto, S. High-resolution image reconstruction with latent diffusion models from human brain activity. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 14453–14463.
6. Massey, J.; et al. Causality, feedback and directed information. In Proceedings of the Proc. Int. Symp. Inf. Theory Applic.(ISITA-90), 1990, Vol. 2, p. 1.
7. Quinn, C.J.; Coleman, T.P.; Kiyavash, N.; Hatsopoulos, N.G. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *Journal of computational neuroscience* **2011**, *30*, 17–44.

 doi:10.20944/preprints202605.1857.v1

8. Wang, Z.; Alahmadi, A.; Zhu, D.C.; Li, T. Causality analysis of fMRI data based on the directed information theory framework. *IEEE Transactions on Biomedical Engineering* **2015**, *63*, 1002–1015.
9. Fraser, A.M.; Swinney, H.L. Independent coordinates for strange attractors from mutual information. *Physical review A* **1986**, *33*, 1134.
10. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. M5 accuracy competition: Results, findings, and conclusions. *International journal of forecasting* **2022**, *38*, 1346–1364.
11. Deco, G.; Jirsa, V.K.; McIntosh, A.R. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature reviews neuroscience* **2011**, *12*, 43–56.
12. Friston, K. The free-energy principle: a unified brain theory? *Nature reviews neuroscience* **2010**, *11*, 127–138.
13. Sankar, L.; Rajagopalan, S.R.; Poor, H.V. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security* **2013**, *8*, 838–852.
14. Allen, E.J.; St-Yves, G.; Wu, Y.; Breedlove, J.L.; Prince, J.S.; Dowdle, L.T.; Nau, M.; Caron, B.; Pestilli, F.; Charest, I.; et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience* **2022**, *25*, 116–126.
15. Wang, L.; Mruczek, R.E.; Arcaro, M.J.; Kastner, S. Probabilistic maps of visual topography in human cortex. *Cerebral cortex* **2015**, *25*, 3911–3931.
16. Henriksson, L.; Karvonen, J.; Salminen-Vaparanta, N.; Railo, H.; Vanni, S. Retinotopic maps, spatial tuning, and locations of human visual areas in surface coordinates characterized with multifocal and blocked FMRI designs. *PloS one* **2012**, *7*, e36859.
17. Soyuhos, O.; Scarpa, A.; Baldauf, D. Distinct Resting-State Connectomes for Face and Scene Perception Predict Individual Task Performance. *Human Brain Mapping* **2026**, *47*, e70498.
18. Johnson, M.R.; Johnson, M.K. Top–down enhancement and suppression of activity in category-selective extrastriate cortex from an act of reflective attention. *Journal of Cognitive Neuroscience* **2009**, *21*, 2320–2327.
19. Avants, B.B.; Epstein, C.L.; Grossman, M.; Gee, J.C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis* **2008**, *12*, 26–41.
20. Cover, T.M. *Elements of information theory*; John Wiley & Sons, 1999.
21. Kvålseth, T.O. On normalized mutual information: measure derivations and properties. *Entropy* **2017**, *19*, 631.
22. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67.
23. Holt, C.C. Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting* **2004**, *20*, 5–10.
24. Graves, A. Long short-term memory. *Supervised sequence labelling with recurrent neural networks* **2012**, pp. 37–45.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
26. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* **2017**.
27. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **2019**, *32*.
28. Huber, P.J. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*; Springer, 1992; pp. 492–518.
29. Kramer, G. *Directed information for channels with feedback*; Vol. 11, Hartung-Gorre Germany, 1998.