GLMsingle: a toolbox for improving single-trial fMRI response estimates

Jacob S. Prince [1*], Ian Charest [2,3], Jan W. Kurzawski [4], John A. Pyles [5], Michael J. Tarr [6], and Kendrick N. Kay [7]

[1]*Department of Psychology, Harvard University, Cambridge, MA, USA*
[2]*Center for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, UK*
[3]*cerebrUM, Département de Psychologie, Université de Montréal, Montréal, Canada*
[4]*Department of Psychology, New York University, New York, NY, USA*
[5]*Center for Human Neuroscience, Department of Psychology, University of Washington, Seattle, WA, USA*
[6]*Department of Psychology, Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA*
[7]*Center for Magnetic Resonance Research (CMRR), Department of Radiology, University of Minnesota, Minneapolis, MN, USA*

* Corresponding author (jacob.samuel.prince@gmail.com)

ABSTRACT

Advances in modern artificial intelligence (AI) have inspired a paradigm shift in human neuroscience, yielding large-scale functional magnetic resonance imaging (fMRI) datasets that provide high-resolution brain responses to tens of thousands of naturalistic visual stimuli. Because such experiments necessarily involve brief stimulus durations and few repetitions of each stimulus, achieving sufficient signal-to-noise ratio can be a major challenge. We address this challenge by introducing *GLMsingle*, a scalable, user-friendly toolbox available in MATLAB and Python that enables accurate estimation of single-trial fMRI responses (glmsingle.org). Requiring only fMRI time-series data and a design matrix as inputs, GLMsingle integrates three techniques for improving the accuracy of trial-wise general linear model (GLM) beta estimates. First, for each voxel, a custom hemodynamic response function (HRF) is identified from a library of candidate functions. Second, cross-validation is used to derive a set of noise regressors from voxels unrelated to the experimental paradigm. Third, to improve the stability of beta estimates for closely spaced trials, betas are regularized on a voxel-wise basis using ridge regression. Applying GLMsingle to the Natural Scenes Dataset and BOLD5000, we find that GLMsingle substantially improves the reliability of beta estimates across visually-responsive cortex in all subjects. Furthermore, these improvements translate into tangible benefits for higher-level analyses relevant to systems and cognitive neuroscience. Specifically, we demonstrate that GLMsingle: (i) improves the decorrelation of response estimates between trials that are nearby in time; (ii) enhances representational similarity between subjects both within and across datasets; and (iii) boosts one-versus-many decoding of visual stimuli. GLMsingle is a publicly available tool that can significantly improve the quality of past, present, and future neuroimaging datasets that sample brain activity across many experimental conditions.

Keywords: fMRI preprocessing, GLM, large-scale datasets, denoising, voxel reliability

INTRODUCTION

Across many scientific disciplines, datasets are rapidly increasing in size and scope. These resources have kickstarted a new era of data-driven scientific discovery (Richards et al., 2019; Jumper et al., 2021; Iten et al., 2020; Ravuri et al., 2021; Schawinski et al., 2018; D'Isanto and Polsterer, 2018). In visual neuroscience, recent efforts to sample individual brains at unprecedented scale and depth have yielded high-resolution functional magnetic resonance imaging (fMRI) datasets in which subjects view thousands of distinct images over several dozen hours of scanning (see Naselaris et al., 2021 for a review). These exciting "condition-rich" datasets are large enough to propel the development of computational models of how humans process complex naturalistic stimuli. For example, resources such as the Natural Scenes Dataset (NSD, Allen et al., 2022), BOLD5000 (Chang et al., 2019), and THINGS (Hebart et al., 2019) may be useful for advancing our ability to characterize the tuning (Bao et al., 2020; Li and Bonner, 2021; Long et al., 2018; Kriegeskorte and Wei, 2021; Popham et al., 2021), topography (Blauch et al., 2021; Doshi and Konkle, 2021; Zhang et al., 2021; Lee et al., 2020), and computations (Yamins et al., 2014; DiCarlo et al., 2012; Freeman et al., 2013; Marques et al., 2021; Horikawa and Kamitani, 2017) performed in visual cortex.

The potential of large-scale datasets to reveal general principles of neural function depends critically on signal-to-noise ratio (SNR), which refers to one's ability to reliably measure distinct neural signatures associated with different stimuli or experimental conditions. Diverse sources of noise affect fMRI data, and these noise sources limit the robustness and interpretability of data analyses (Liu, 2016; Kay et al., 2013). For example, subject head motion, scanner instabilities, physiological noise, and thermal noise all contribute unwanted variability to fMRI data. Noise is especially problematic in studies that sample a large number of conditions, since the number of repetitions of each condition is typically limited, resulting in noisy responses even after trial-averaging.

The approach we have developed to mitigate the effects of noise comes in the context of general linear model (GLM) analysis of fMRI time-series data (Dale, 1999; Monti, 2011). We assume that the goal of the GLM analysis is to estimate beta weights representing the blood oxygenation level dependent (BOLD) response amplitude evoked by different experimental conditions. In this context, we define *noise* as variability observed across repeated instances of a given condition. Therefore, methods that decrease such variability are desirable. Our approach seeks to maximize data quality at the level of individual voxels in individual subjects (as opposed to data quality assessed only at the region or group level), and seeks to obtain response estimates for single trials. These desiderata are powerful; if achieved, they can flexibly support a wide range of subsequent analyses including relating brain responses to trial-wise behavioral measures and pooling data across trials, brain regions, and/or subjects.

To realize these goals, we introduce *GLMsingle*, a user-friendly software toolbox (with both MATLAB and Python implementations) that performs single-trial BOLD response estimation. Given fMRI time-series data and a design matrix indicating the onsets of experimental conditions, GLMsingle implements a set of optimizations that target three aspects of the GLM framework (**Figure 1**):

1. The choice of hemodynamic response function (HRF) to convolve with the design matrix
2. The inclusion of nuisance regressors that account for components of the data that are thought to be noise
3. The use of regularization to improve the accuracy of the final beta estimates

Importantly, to enable fluid application to even the largest fMRI datasets, GLMsingle is fully automated (no manual setting of parameters) and can be executed efficiently even when gigabytes of fMRI data are passed as input.

We previously used the GLMsingle algorithm to estimate BOLD responses in the NSD dataset (Allen et al., 2022). While the optimizations implemented in GLMsingle had a positive impact on data quality, it was not apparent whether the improvements would generalize to other datasets. The goal of this paper is to provide a standalone description of GLMsingle and to rigorously assess performance not only on NSD, but also on BOLD5000 (Chang et al., 2019), a distinct fMRI dataset acquired with different subjects, at different field strength, and with a different experimental design (see *Methods*). In both datasets, we show that the optimizations implemented in GLMsingle dramatically improve the reliability of GLM beta estimates. We also study the effect of these optimizations on downstream analyses that are of particular relevance to systems and cognitive neuroscience, including representational similarity

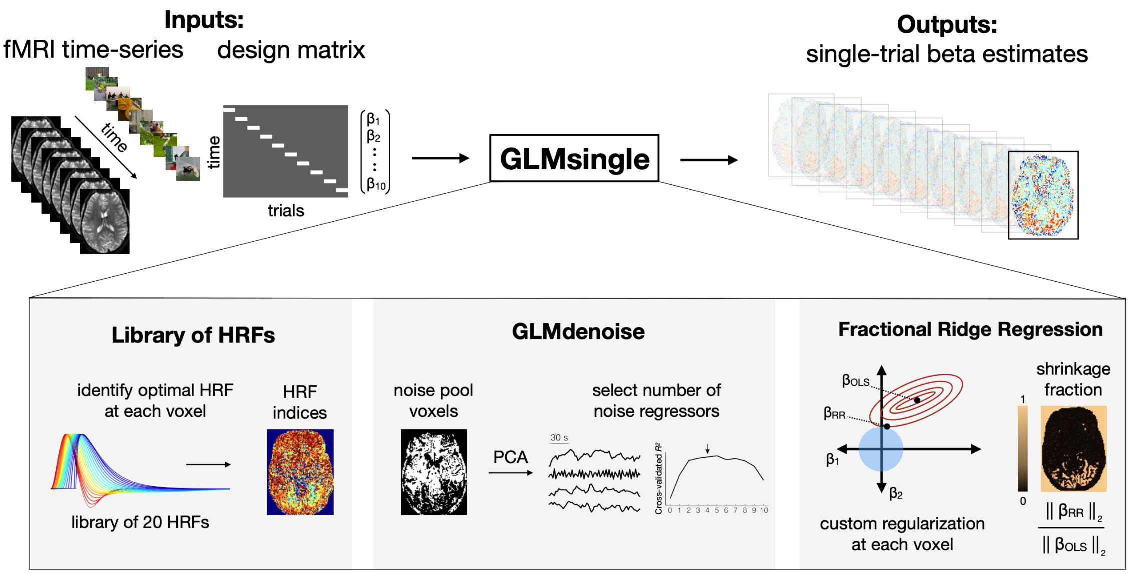

Figure 1: Overview of GLMsingle
GLMsingle takes as input a design matrix (where each column indicates the onset times for a given condition) and fMRI time-series in either volumetric or surface space, and returns as output an estimate of single-trial BOLD response amplitudes (beta weights). GLMsingle incorporates three techniques designed to optimize the quality of beta estimates: first, the use of a library of hemodynamic response functions (HRFs), where the best-fitting HRF from the library is chosen for each voxel; second, an adaptation of GLMdenoise (Kay et al., 2013) to the single-trial GLM framework, where data-derived nuisance regressors are identified and used to remove noise from beta estimates; and third, an efficient re-parameterization of ridge regression (Rokem and Kay, 2020) as a method for dampening the noise inflation caused by correlated single-trial GLM predictors.

analysis (RSA) (Kriegeskorte et al., 2008) and multivoxel pattern analysis (MVPA) (Haxby et al., 2001, Norman et al., 2006, Poldrack et al., 2011). In all analyses, we observe improvements in key outcome metrics, suggesting that GLMsingle meaningfully improves the ability of researchers to gain insight into neural representation and computation. Our findings demonstrate that GLMsingle affords the neuroimaging community a clear opportunity for improved data quality. Online materials (code, documentation, example scripts) pertaining to GLMsingle are available at glmsingle.org.

RESULTS

To assess the impact of GLMsingle, we evaluate four different types of single-trial response estimates (henceforth, *beta versions*). The first arises from a baseline procedure that reflects a typical GLM approach for fMRI analysis (beta version $b1$), and each subsequent beta version ($b2$-$b4$) incorporates an additional strategy for optimizing model fits and mitigating the effects of noise. The final beta version ($b4$) contains the complete set of optimizations provided by the GLMsingle toolbox. The GLMsingle algorithm consists of the following steps:

1. A baseline single-trial GLM is used to model each stimulus trial separately using a canonical HRF. This provides a useful baseline for comparison ($b1$**: AssumeHRF**).

2. An optimal HRF is identified for each voxel (Allen et al., 2022) by iteratively fitting a set of GLMs, each time using a different HRF from a library of 20 HRFs. For each voxel, we

identify the HRF that provides the best fit to the data (highest variance explained), and inherit the single-trial betas associated with that HRF ($b2$**: FitHRF**).

3. GLMdenoise (Kay et al., 2013; Charest et al., 2018) is used to determine nuisance regressors to include in the model. Principal components analysis is applied to time-series data from a pool of noise voxels (see *Methods* for details), and the top principal components are added one at a time to the GLM until cross-validated variance explained is maximized on-average across voxels ($b3$**: FitHRF + GLMdenoise**).

4. With the nuisance regressors determined, fractional ridge regression (Rokem and Kay, 2020) is used to regularize the single-trial betas, using a custom amount of regularization for each voxel, determined via cross-validation ($b4$**: FitHRF + GLMdenoise + RR**).

GLMsingle improves the reliability of beta estimates

We first examine the effect of GLMsingle on the test-retest reliability of voxels across relevant regions of visual cortex in NSD and BOLD5000 (**Figure 2**). Our reliability procedure measures the consistency of a voxel's response profile (using Pearson r) over repeated presentations of the same stimuli, revealing areas of the brain containing stable BOLD responses. This straightforward approach enables direct comparison of data quality between different beta versions.

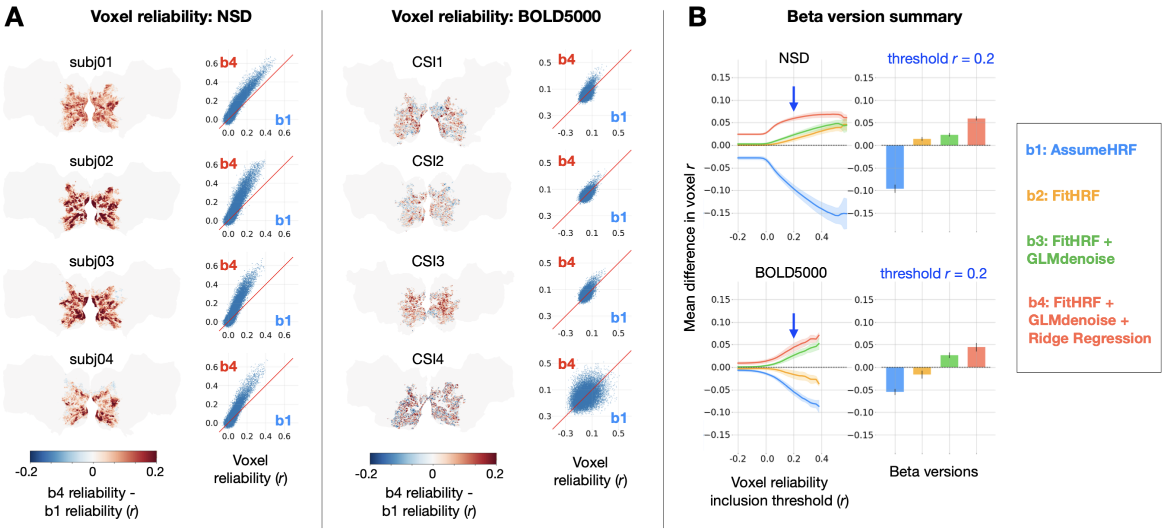

Figure 2: Impact of GLMsingle on voxel test-retest reliability

To compute reliability for a given voxel, we measure the test-retest Pearson correlation of GLM beta profiles over repeated presentations of the same stimuli (see Methods). (A) Differences in reliability between $b1$ *(derived from a baseline GLM) and* $b4$ *(the final output of GLMsingle) are plotted within a liberal mask of visual cortex (nsdgeneral ROI). Scatter plots show reliability values for individual voxels. (B) Relative differences in mean reliability within the nsdgeneral ROI. For each voxel, we computed the mean reliability value over all beta versions being considered (*$b1$*-*$b4$*), and then used this as the basis for thresholding voxels (from Pearson* $r = -0.2$ *to 0.6). At each threshold level, for each beta version, we compute the voxel-wise difference between the reliability of that specific beta version and the mean reliability value, and then average these difference values across voxels within the nsdgeneral ROI. The traces in the first column indicate the mean (+/- SEM) across subjects within each dataset. The bars in the second column indicate subject-averaged differences in reliability at threshold* $r = 0.2$*. The relative improvement in reliability due to GLMsingle (*$b1$ *vs.* $b4$*) tends to increase when examining voxels with higher reliability, and each optimization stage within GLMsingle (HRF fitting, GLMdenoise, ridge regression) confers added benefit to voxel reliability.*

We directly compared the $b1$ and $b4$ beta versions for each subject within a liberal mask of visual cortex (nsdgeneral ROI), finding widespread increases in reliability when comparing GLMsingle to baseline

(**Figure 2a**). The positive effect is nearly uniform across voxels in NSD. In BOLD5000, as in NSD, we see aggregate benefits when comparing $b1$ and $b4$, though results for individual voxels are more variable. A likely explanation for this is that reliability metrics are inherently noisier due to the smaller number of repeated stimuli in BOLD5000.

To summarize the impact of GLMsingle in NSD and BOLD5000, we compared the performance of $b1$-$b4$ for individual subjects, across different voxel reliability thresholds (**Figure 2b**). We find that all subjects show clear improvement from $b1$ to $b4$ and the improvement in reliability due to GLMsingle tends to increase when examining voxels that respond more reliably to experimental stimuli. Furthermore, examining reliability in intermediate beta versions ($b2$ and $b3$) – which implement HRF optimization and GLMdenoise, respectively – reveals that each successive stage of processing in GLMsingle tends to confer added benefit to voxel reliability compared to baseline ($b1$).

We next compared GLMsingle to Least-Squares Separate (LSS), a popular technique for robust signal estimation in rapid event-related designs (Mumford et al., 2012, 2014; Abdulrahman and Henson, 2016). The LSS procedure fits a separate GLM for each stimulus, where the trial of interest is modeled as one regressor, and all other (non-target) trials are collapsed into a second regressor. LSS provides a useful point of comparison for ridge regression, as both strategies seek to mitigate the instabilities in GLM estimation that can arise from having correlated single-trial predictors. To directly compare GLMsingle to LSS, we computed auxiliary GLMsingle beta versions that do not incorporate GLMdenoise. This allows us to isolate the effect of the GLM estimation procedure (i.e., LSS vs. fractional ridge regression).

For both the case of an assumed HRF and the case of voxel-wise tailored HRFs, we find that fractional ridge regression yields more reliable signal estimates than LSS (**Figure 3**). These improvements are most pronounced in the most reliable voxels (**Figure 3c**). LSS can be viewed as applying heavy regularization uniformly across voxels, while our ridge regression approach is more flexible, tailoring the degree of regularization to the SNR of each voxel. Heavy regularization may actually degrade the quality of signal estimates in reliable voxels, and our approach avoids this possibility.

We then performed a complete assessment of all auxiliary beta versions and the primary versions ($b1$-$b4$), in order to determine whether any other analysis strategy could achieve parity with $b4$ in the quality of GLM outputs. Reassuringly, when summarizing the relative quality of all 8 beta versions over a range of reliability thresholds, we observe superior performance from $b4$, the default output of GLMsingle **(Figure 3a)**.

GLMsingle relies on an internal cross-validation procedure through which key hyperparameters (the number of noise regressors and the voxel-wise levels of ridge regression regularization) are optimized to maximize the consistency of responses across condition repetitions. This raises a possible concern that our reliability estimates (e.g. **Figure 2**) are somewhat optimistic. As a strict assessment of reliability, we repeated the reliability quantification for each of the 8 beta versions, this time computing test-retest correlation values using only beta responses obtained from completely separate data partitions. We find that results are broadly unchanged using this more stringent evaluation procedure **(Figure 3b).**

GLMsingle helps disentangle neural responses to neighboring trials

Thus far, we have established that GLMsingle provides BOLD response estimates that have substantially improved reliability compared to a baseline GLM. In the remainder of this paper, we explore whether these improvements have tangible consequences for downstream analyses relevant for cognitive and systems neuroscience. We first examine whether GLMsingle is able to more effectively disentangle neural responses to proximal stimuli, as inaccurate single-trial GLM estimation may manifest as high similarity (temporal autocorrelation) between beta maps from nearby trials. We computed dataset-

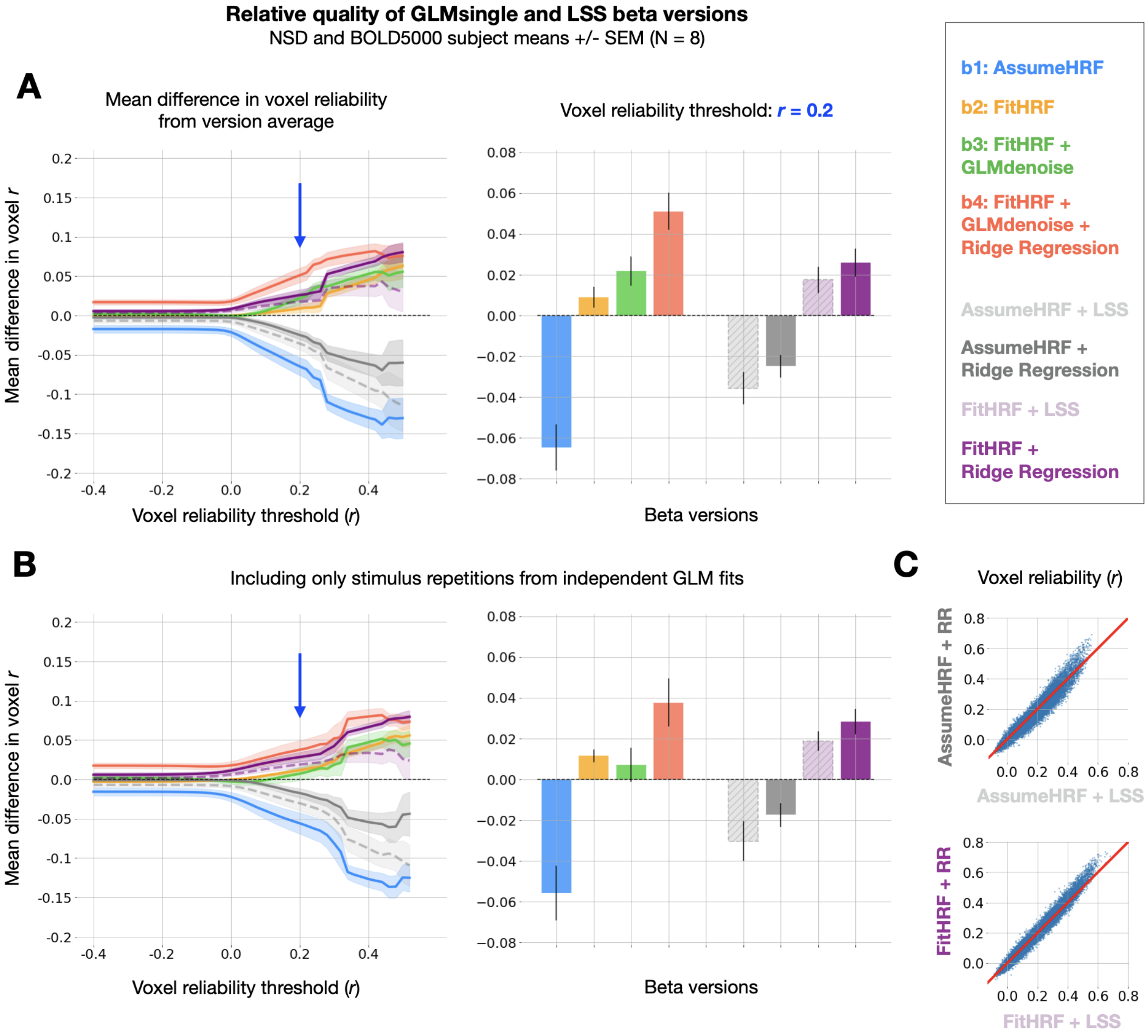

Figure 3: Comparison between GLMsingle and LSS

(A) Left panel: relative differences in mean reliability between beta versions. 8 beta versions are compared: b1-b4, and the 4 auxiliary beta versions used to compare GLMsingle and Least-Squares Separate (LSS). LSS betas (dashed traces) are compared to those estimated using fractional ridge regression (RR, solid traces), when using a canonical HRF (LSS, light gray vs. RR, dark gray) and when performing HRF optimization (LSS, light purple vs. RR, dark purple). Right panel: Summary of performance at threshold level $r = 0.2$. Error bars reflect the standard error of the mean, computed over the 8 subjects analyzed from NSD and BOLD5000. Fractional ridge regression yields more reliable signal estimates than LSS across voxel reliability levels. (B) Same as Panel (A), except that reliability computations occur only between image repetitions processed in independent partitions of fMRI data. Qualitative patterns are unchanged. (C) Scatter plots comparing voxel reliability between corresponding LSS and GLMsingle beta versions (top: AssumeHRF; bottom: FitHRF). We show results for an example subject (NSD subj01, nsdgeneral ROI). The advantage of ridge regression over LSS is most apparent in the most reliable voxels.

averaged temporal similarity matrices, revealing the degree of temporal autocorrelation in each beta version (**Figure 4**). Temporal autocorrelation manifests as non-zero correlation values off the diagonal of the temporal similarity matrices, and is presumably undesirable.

In a baseline GLM that uses a canonical HRF and ordinary least squares (OLS) fitting ($b1$), we observe striking patterns of temporal autocorrelation extending several dozen trials forward in time. This is true in both NSD, which has a rapid event-related design (a new stimulus presented every 4 s),

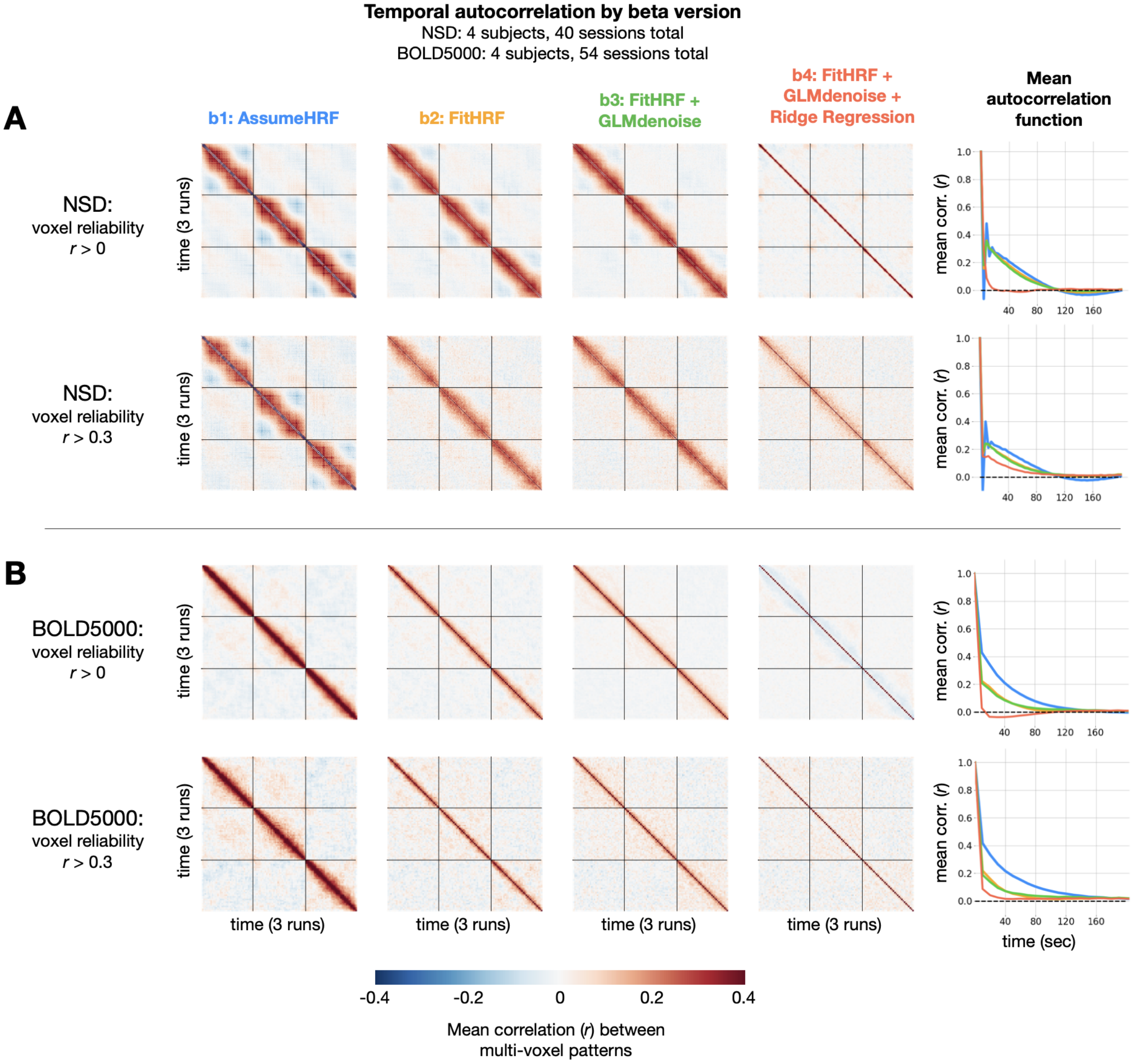

Figure 4: Impact of GLMsingle on temporal autocorrelation

For each dataset, we compute the degree of temporal autocorrelation in each beta version by averaging session-wise representational similarity matrices over subjects. We plot results arising from analysis of voxels at two different reliability thresholds ($r = 0$ and $r = 0.3$) for NSD (A) and BOLD5000 (B). Assuming that ground-truth neural responses to consecutive trials should be uncorrelated on average, positive (or negative) Pearson r values off the diagonal imply sub-optimal estimation of BOLD responses. In the right-most column, we plot mean autocorrelation between all pairs of timepoints. Applying GLMsingle ($b4$) results in a substantial decrease in temporal autocorrelation compared to a baseline GLM approach ($b1$).

as well as in BOLD5000, where stimuli are spaced 10 s apart to alleviate issues relating to signal overlap. To quantify these effects, we compute mean temporal autocorrelation as a function of time post-stimulus for each beta version. In NSD, for the baseline GLM ($b1$), positive correlations are as high as $r = 0.5$ for consecutive trials, and gradually reduce to around $r = 0$ after around 100 s (**Figure 4a**). In BOLD5000, $b1$ autocorrelation peaks as high as around $r = 0.4$ for consecutive trials, requiring nearly 160 s to reduce to $r = 0$ (**Figure 4b**). We speculate that the relatively long timescale of the correlations reflects the long timescale of hemodynamic responses (the post-undershoot can extend for 30 s or longer) and/or the slow nature of (low-frequency) physiological noise related to cardiac and respiratory variation. Notably, mean beta maps from successive trials in NSD are *anticorrelated* for $b1$, a known artifact of OLS fitting in the case of high multicollinearity between GLM predictors (Mumford et al., 2014; Soch et al., 2020).

When applying GLMsingle, these patterns of temporal autocorrelation change dramatically. In NSD $b4$, autocorrelation drops to $r = 0$ much more rapidly than in $b1$, and in BOLD5000, beta maps from successive trials in $b4$ are now nearly uncorrelated on average. This is an expected outcome, since the stimuli in NSD and BOLD5000 are ordered pseudorandomly. In both datasets, an intermediate beta version ($b2$) containing only HRF optimization confers marginal benefit over $b1$, but the most dramatic improvements come from the addition of both GLMdenoise and fractional ridge regression ($b4$). Overall, these results demonstrate the utility of GLMsingle for disentangling neural responses to nearby stimuli in event-related designs, even when events are presented relatively slowly (as in BOLD5000).

GLMsingle improves between-subject representational similarity across datasets

Large-scale datasets such as NSD and BOLD5000 are well-suited for representational analyses (e.g., RSA) that compare evoked neural response patterns between individual subjects, across different experimental modalities, and against computational models (e.g., deep neural networks, see Kriegeskorte, 2015, Serre, 2019 for review.) In almost all such studies, representational analyses presume that the same set of stimuli will evoke reasonably similar responses across subjects. As such, given the ubiquity of noise in fMRI, it is reasonable to expect that improving the accuracy of single-trial response estimates should yield representations that are more similar across individuals.

To compare representations between subjects, we used the approach of RSA (Kriegeskorte et al., 2008). First, we isolated stimuli that overlap between BOLD5000 and the subset of NSD analyzed for this manuscript (the first 10 sessions from each subject). Using these 241 stimuli, we constructed representational dissimilarity matrices (RDMs) using repetition-averaged betas from each individual, and then correlated all pairs of subject RDMs within and between datasets. Note that GLMsingle is not designed to enhance or optimize cross-subject representational similarity; as such, it is informative to examine RSA correlations between subjects as a way of assessing methods for denoising (Charest et al., 2018). Strikingly, in comparing beta versions $b1$ and $b4$, we observe a consistent strengthening of RDM correspondence **(Figure 5b)**. This trend held within NSD, within BOLD5000, and when comparing the RDMs of subject pairs between the two datasets. The latter result is especially striking given the many methodological differences between NSD and BOLD5000: fMRI data were collected at different sites on different scanners, at different field strengths (7T vs. 3T), with different behavioral tasks, and with different inter-stimulus intervals (4 s vs. 10 s).

These results indicate that GLMsingle, through its multifaceted approach to mitigating the effects of noise, helps reveal meaningful shared variance in neural responses across individuals who viewed the same stimuli. The GLMsingle toolbox may therefore be a key resource for future fMRI studies seeking to stitch together data across subjects from different sites or cohorts.

GLMsingle enables fine-grained image-level MVPA decoding

As a final analysis, we assessed the effect of GLMsingle on the results of multivoxel pattern analysis (MVPA). In a "one-vs.-many" classification paradigm, we trained linear SVM models for each subject to predict image identity from neural response patterns. The baseline GLM ($b1$) classification accuracy was slightly above chance on average for the subjects in NSD and BOLD5000 when including all visual cortex voxels (**Figure 6a**, blue traces). Performing the same MVPA procedure using GLMsingle betas ($b4$), we observe that mean accuracy approximately triples in NSD and doubles in BOLD5000 (**Figure 6a**, red traces). Moreover, in both datasets we observe a substantial increase in classification accuracies

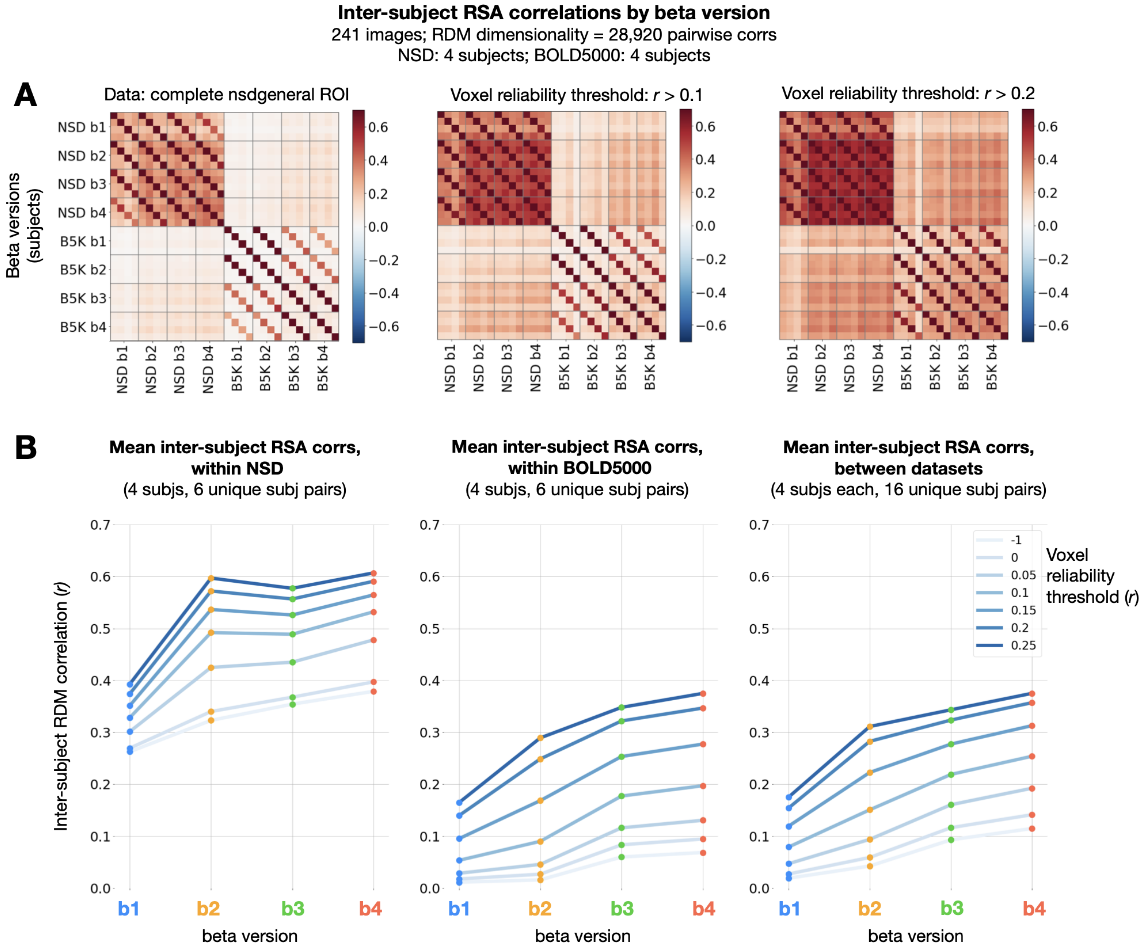

Figure 5: Impact of GLMsingle on inter-subject RSA correlations
(A) Correlations of RDMs across all pairs of subjects and beta versions, at 3 different voxel reliability thresholds. We compute RDMs for each subject and beta version using Pearson dissimilarity (1 - r) over repetition-averaged betas within the nsdgeneral ROI. Grid lines separate beta versions from one another, an individual cell reflects the RDM correlation between one pair of subjects, and cross-dataset comparisons occupy the top-right and bottom-left quadrants of the matrices. (B) Mean inter-subject RDMs correlations within NSD (left), within BOLD5000 (center), and between the two datasets (right). GLMsingle (b4) yields a considerable strengthening of RDM correspondence for each subject pair being considered, within and between datasets.

with increasing voxel reliability threshold, with the most dramatic improvements achieved using $b4$ in NSD (**Figure 6a**, left panel, right-most bins).

The level of performance that GLMsingle facilitates on this challenging multi-way decoding task highlights the ability of the technique to accurately identify and model the stable structure contained in noisy fMRI time-series. To illustrate this point, we performed 2D multidimensional scaling (MDS, Borg and Groenen, 2005) using NSD betas that were included in MVPA. Comparing results between beta versions $b1$ and $b4$, we observe improved clarity of an animacy division in the representational space of an example subject (**Figure 6b**).

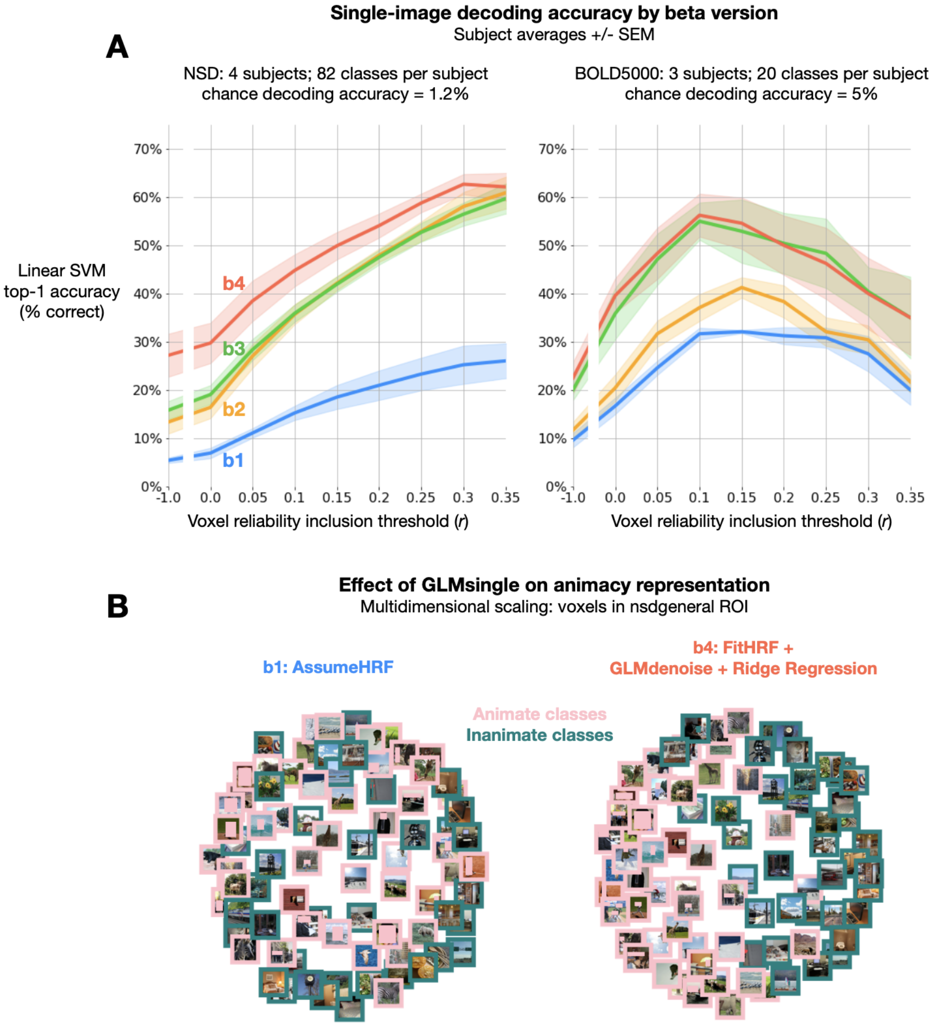

Figure 6: Impact of GLMsingle on MVPA decoding accuracy
(A) Image-level linear SVM decoding accuracy by beta version. At each reliability threshold, we compute the mean decoding accuracy over subjects within each dataset, as well as the standard error of the mean. Classifiers are trained on $n - 1$ available image repetitions, and tested on the held-out repetition, with accuracy averaged over cross-validation folds. Applying GLMsingle ($b4$) yields dramatic increases in image decodability compared to a baseline GLM ($b1$). (B) The effect of GLMsingle on animacy representation is shown in an example NSD subject (subj01) using multi-dimensional scaling. GLMsingle clarifies the division in representational space between stimuli containing animate and inanimate objects. COCO stimuli containing identifiable human faces are masked with a rectangle for the sake of privacy.

DISCUSSION

As scientific datasets grow in scale and scope, new techniques for data processing will help to unlock their potential. This is especially true in human neuroscience where data remain both expensive and time-consuming to collect (Naselaris et al., 2021). This paper has introduced GLMsingle, a publicly available toolbox for analyzing fMRI time-series data that leverages data-driven techniques to improve the accuracy of single-trial fMRI response estimates. We have tested GLMsingle extensively using NSD and BOLD5000, two of the largest fMRI datasets that densely sample responses within individuals. For both datasets, analyses of the response estimates provided by GLMsingle indicate substantial improvements in several key metrics of interest to neuroscientists: (i) enhanced test-retest reliability of voxel response profiles, a straightforward metric of data quality; (ii) reduced temporal autocorrelation, a common fMRI effect that is presumably undesirable and especially prominent in rapid event-related designs; (iii) increased representational similarity across subjects both within and across datasets; and (iv) improved multivariate pattern classification performance when discriminating responses evoked by individual images.

Principles underlying GLMsingle

GLMsingle incorporates three optimization procedures to improve the estimation of fMRI responses:

1. *HRF fitting.* GLMsingle uses a "library of HRFs" technique to select the most appropriate HRF to use for each voxel in a given dataset (Allen et al., 2022). This library consists of a set of 20 HRFs that were derived from experimental data (specifically, the first NSD scan session acquired in each of the 8 NSD subjects). It is well known that variations in HRFs exist across voxels, brain areas, and subjects, and that mismodeling the timecourse of a voxel may lead to suboptimal analysis outcomes (Handwerker et al., 2004, 2012). Imposing constraints on HRF selection by choosing from a fixed set of HRFs avoids the instability (high variance) associated with more flexible timecourse modeling approaches, such as finite impulse response modeling (Kay et al., 2008; Bai and Kantor, 2007). Variations in timecourse shapes in the HRF library reflect a continuum between short-delay, narrow-width timecourses to long-delay, broad-width timecourses, and are likely caused by variations in the contribution of large vessels to the BOLD response observed in a voxel (Kay et al., 2020).

2. *Data-driven denoising.* Incorporating an adaptation of the GLMdenoise technique (Kay et al., 2013), GLMsingle uses principal components analysis to calculate potential nuisance regressors from fMRI time-series data observed in voxels that are deemed unrelated to the experimental paradigm. These regressors are incorporated into the GLM using a cross-validation procedure to determine the optimal number of nuisance regressors to add. A key aspect of our approach is the acknowledgement that including increasing numbers of nuisance regressors will, at some point, cause overfitting and degradation of results (Kay et al., 2013); this motivates the use of cross-validation to determine the optimal level of model complexity.

3. *Regularization of GLM weights.* To improve the accuracy of single-trial GLM response estimates, GLMsingle uses fractional ridge regression (Rokem and Kay, 2020), with an optimal degree of regularization identified for each voxel, again using cross-validation. The improvements afforded by this procedure are due to the substantial amount of overlap of the fMRI response across successive trials, unless very long (> 30 s) inter-stimulus intervals are used. It is well known that, in the context of ordinary least squares estimation, two predictors that are correlated (or anti-correlated) will have reduced estimation precision compared to the scenario in which the predictors are uncorrelated (Mumford et al., 2012; Soch et al., 2020). For rapid event-related designs, predictors for consecutive trials are typically correlated, and ordinary least-squares estimates will suffer from high levels of instability. Ridge regression imposes a shrinkage prior (penalizing the sum of the squares of the beta estimates), which can, in principle, dampen the effects of noise and improve out-of-sample generalizability of the beta estimates.

Ideal use-cases for GLMsingle

GLMsingle is designed to be general in its application. It uses data-driven procedures that automatically adapt to the signal-to-noise characteristics of a given dataset. For example, in datasets where structured noise is prevalent, appropriate nuisance regressors will automatically be included, whereas in datasets with very little structured noise (e.g., low head motion), fewer (or no) nuisance regressors will be included. As another example, for experimental designs with high temporal overlap between consecutive trials or high levels of noise, relatively strong levels of shrinkage regularization will likely be selected.

GLMsingle is a general technique that can be fruitfully applied to nearly *any* fMRI experiment involving discrete events (including block designs). However, we recognize that integrating a new tool into an analysis workflow requires effort. Therefore, we anticipate that the most consequential impact of GLMsingle will be observed for study designs with low sensitivity (such as condition-rich designs).

Potential limitations to consider when applying GLMsingle

GLMsingle relies on cross-validation to determine two key hyperparameters: (i) the number of nuisance regressors to use in the GLM as derived by applying PCA to data from the noise pool voxels; and (ii) the amount of ridge-regression shrinkage to apply for each voxel. Although the data-driven nature of the technique is one of its strengths (since it adapts to the characteristics of each dataset), it is also a potential limitation. First, a prerequisite for application of GLMsingle is the existence of at least some repeated trials in a given dataset. A dataset consisting only of experimental conditions with a single occurrence each cannot be used in conjunction with the cross-validated procedures for determining the optimal number of nuisance regressors and the voxel shrinkage fractions. Second, since data are invariably noisy, the determination of hyperparameters is subject to noise, and it is not guaranteed that hyperparameter estimates will be accurate in all possible data situations. It remains an open question for further investigation what the minimum data requirements are for reasonably accurate hyperparameter estimation.

Given the requirement of repeated discrete events, GLMsingle is not applicable to resting-state data, since they contain no explicit task structure. Similarly, GLMsingle is not suitable for experiments that involve continuous event structures – for example, movie watching, storytelling, dynamic exploration, game-playing — unless certain events within the task are coded as discrete, repeated instances. For example, the appearance on-screen of a particular character could be treated as a repeated "event" in constructing a design matrix. Or, as another example, certain words or parts of speech could be treated as "events" within a continuous auditory or linguistic experiment.

It is important to consider whether denoising comes at the potential cost of introducing bias (Kay, 2022). Considering each component of GLMsingle, we believe that the risk of bias is minimal for most use cases. First, considering the library-of-HRFs approach, we note that the conventional approach of using a fixed canonical HRF actually incurs more risk of biasing response estimates than does an approach that attempts to flexibly capture variations in HRFs. Nonetheless, we acknowledge that the library may not necessarily capture all HRF shapes, and this represents one possible source of bias (though likely minor). Second, considering the GLMdenoise procedure, we note that data-derived nuisance regressors are not blindly removed from the time-series data prior to modeling, as this would pose a clear risk of removing experimentally-driven signals, thereby leading to bias (Liu et al., 2001). Rather, by including both task-related regressors and nuisance regressors in the GLM, the model can appropriately partition variance between signal and noise sources. Third, considering ridge regression, we note that shrinkage can be viewed as a form of temporal smoothing, in the sense that beta weights from temporally adjacent trials are biased to be more similar in magnitude. While this is indeed a source of bias, this should be concerning only for investigations where relative responses for nearby trials are of specific interest (e.g., studies of repetition suppression). For other investigations, and especially for experiments where condition ordering is pseudorandom, it is unlikely that this form of temporal regularization and its associated bias would lead to incorrect scientific inferences.

Online example scripts and tutorials

To enable easy adoption of GLMsingle, we provide extensive documentation and example scripts for common neuroimaging use-cases (glmsingle.org). Publicly available online resources include code implementation of GLMsingle in both MATLAB and Python, example scripts and notebooks, technical documentation, and answers to frequently asked questions. The GLMsingle pipeline is designed to be easy to implement in different neuroimaging pipelines. The example scripts we provide illustrate typical GLMsingle usage for both event-related and block designs. These scripts guide the user through basic calls to GLMsingle, using representative, small-scale example datasets. We hope these practical resources facilitate the application of GLMsingle to existing and future neuroimaging datasets.

Conclusion

Our results suggest that GLMsingle represents a methodological advancement that will help improve data quality across different fMRI designs. While improvements in MR hardware (e.g. magnetic field strength, RF coil, pulse sequences) and experimental design (e.g. optimized study design and trial distributions) may contribute to improved data quality, the benefits of GLMsingle demonstrated in this paper make clear that data processing techniques are another critical factor that can profoundly impact SNR and overall experimental power. As an analogy, we observe that the rapid (and annual) improvement in cell phone cameras has been driven in large part by advances in image analysis algorithms. As summarized by an Apple executive, "[while sensor quality has improved], increasingly, what makes incredible photos possible aren't just the sensor and the lens but the chip and the software that runs on it" (Wilson, 2018). We suggest that GLMsingle represents a similar advance in signal processing for fMRI.

MATERIALS AND METHODS

Description of GLMsingle

Inputs to GLMsingle

GLMsingle expects that input fMRI data have been preprocessed with motion correction at minimum, and ideally slice time correction as well. Additional common preprocessing steps such as compensation for spatial distortion, spatial smoothing, or registration to an anatomical space (or atlas space) are all compatible with GLMsingle without any complications. Detrending or high-pass filtering the time-series data is not necessary, as low-frequency fluctuations are modeled as part of the GLM fitting procedure. The input fMRI data can be supplied in either volumetric or surface format. Besides fMRI data, the other user-provided input to GLMsingle is an array of design matrices corresponding to each run of the time-series data, indicating the sequence of events that occurred during the runs. GLMsingle expects that these are matrices with dimensions (time x conditions), where each column corresponds to a single condition and consists of 0s except for 1s indicating the onset times for that condition. Further details about data formats are provided in the online code repository.

GLMsingle overview

GLMsingle consists of three main analysis components. The first component is the use of a library of hemodynamic response functions (HRFs) to identify the best-fitting HRF for each voxel. This simple approach for compensating for differences in hemodynamic timecourses across voxels (Handwerker et al., 2004) has several appealing features: it invariably provides well-regularized HRF estimates, and it is efficient and can be executed with reasonable computational cost. The second component is an adaptation of GLMdenoise to a single-trial GLM framework. GLMdenoise is a previously introduced technique (Kay et al., 2013) in which data-derived nuisance regressors are identified and used to remove noise from—and therefore improve the accuracy of—beta estimates. The third analysis component is an application of ridge regression (Hoerl and Kennard, 1970) as a method for dampening the noise inflation caused by correlated single-trial GLM predictors. To determine the optimal level of regularization for each voxel, we make use of a recently developed efficient re-parameterization of ridge regression called "fractional ridge regression" (Rokem and Kay, 2020).

Derivation of the library of HRFs

The HRF library incorporated into GLMsingle was previously used for signal estimation in analyzing the Natural Scenes Dataset. Complete details on the derivation procedure for the HRF library can be found in the NSD dataset paper (Allen et al., 2022). In brief, empirically-observed BOLD timecourses were subject to principal components analysis, projected onto the unit sphere, and parameterized using a path consisting of 20 regularly-spaced points through the area of greatest data density. The timecourses corresponding to the resulting set of 20 points were fit using a double-gamma function as implemented in SPM's spm_hrf.m, yielding a fixed library of 20 HRFs. This library is the default in GLMsingle, and was used for all analyses of the NSD and BOLD5000 datasets described here. In future work, it is possible to refine or expand the HRF library (e.g., by deriving it from a larger pool of subjects, or by restricting estimation to individual subjects).

Cross-validation framework for single-trial GLM

The GLMdenoise and ridge regression analysis components of GLMsingle both require tuning of hyperparameters (specifically, the number of nuisance regressors to include in GLM fitting and the regularization level to use for each voxel). To determine the optimal setting of hyperparameters, we use a cross-validation approach in which out-of-sample predictions are generated for single-trial beta estimates. Performing cross-validation on single-trial betas, as opposed to time-series data, simplifies and reduces the computational requirements of the cross-validation procedure. Note that because of cross-validation, although GLMsingle produces estimates of responses to single trials, it does require the existence of and information regarding repeated trials (that is, trials for which the experimental manipulation is the same and expected to produce similar brain responses). This requirement is fairly minimal, as most fMRI experiments are designed in this manner.

The first step of the cross-validation procedure is to analyze all of the available data using a generic GLM. In the case of GLMdenoise, this amounts to the inclusion of zero nuisance regressors; in the case of ridge regression, this amounts to the use of a shrinkage fraction of 1, which corresponds to ordinary least-squares regression. In both cases, the generic analysis produces a full set of unregularized single-trial betas (e.g., in one NSD session, there are 750 single-trial betas distributed across 12 runs, and in one BOLD5000 session, there are either 370 or 333 single-trial betas distributed across either 10 or 9 runs). The second step of the procedure is to perform a grid search over values of the hyperparameter (e.g., number of GLMdenoise nuisance regressors; ridge regression shrinkage fraction). For each value, we assess how well the resulting beta estimates generalize to left-out runs. By default, for all cross-validation procedures, GLMsingle implements the following leave-one-run-out routine: (1) one run is held out as the validation run, and experimental conditions that occur in both the training runs and the validation run are identified; (2) squared errors between the regularized beta estimates from the training runs and the unregularized beta estimates from the validation run are computed; (3) this procedure is repeated iteratively, with each run serving as the validation run, and errors are summed across iterations.

GLMsingle algorithm

Having described the essential aspects of the estimation framework above, we now turn to the steps in the GLMsingle algorithm. GLMsingle involves fitting several different GLM variants. Each variant includes polynomial regressors to characterize the baseline signal level: for each run, we include polynomials of degrees 0 through $round(L/2)$ where L is the duration in minutes of the run.

1. *Fit a simple ON-OFF GLM.* In this model, all trials are treated as instances of a single experimental condition, and a canonical HRF is used. Thus, there is a single "ON-OFF" predictor that attempts to capture signals driven by the experiment. The utility of this simple model is to provide variance explained (R^2) values that help indicate which voxels carry experimentally-driven signals.

2. *Fit a baseline single-trial GLM.* In this model, each stimulus trial is modeled separately using a canonical HRF. This model provides a useful baseline that can be used for comparison against models that incorporate more advanced features (as described below).

3. *Identify an HRF for each voxel.* We fit the data multiple times with a single-trial GLM, each time using a different HRF from the library of HRFs. For each voxel, we identify which HRF provides the best fit to the data (highest variance explained), and inherit the single-trial betas associated with that HRF. Note that the final model for each voxel involves a single chosen HRF from the library.

4. *Use GLMdenoise to determine nuisance regressors to include in the model.* We define a pool of noise voxels (brain voxels that have low ON-OFF R^2, according to an automatically determined threshold) and then perform principal components analysis on the time-series data associated with these voxels (separately for each run). The top principal components (each of which is a timecourse) are added one at a time to the GLM until cross-validation performance is maximized on-average across voxels. The inclusion of these nuisance regressors is intended to capture diverse sources of noise that may be contributing to the time-series data in each voxel.

5. *Use fractional ridge regression to regularize single-trial betas.* With the nuisance regressors determined, we use fractional ridge regression to determine the final estimated single-trial betas. This is done by systematically evaluating different shrinkage fractions. The shrinkage fraction for a given voxel is simply the ratio between the vector length of the set of betas estimated by ridge regression and the vector length of the set of betas returned by ordinary least-squares estimation, and ranges from 0 (maximal regularization) to 1 (no regularization). For each voxel, in the context of a GLM that incorporates the specific HRF chosen for that voxel as well as the identified nuisance regressors, cross-validation is used to select the optimal shrinkage fraction.

The default behavior of GLMsingle is to return beta weights in units of percent signal change by dividing by the mean signal intensity observed at each voxel and multiplying by 100. To preserve the interpretability of GLM betas as percent signal change even after applying shrinkage via ridge regression, we apply a post-hoc scaling and offset on the betas obtained for each given voxel in order to match, in a least-squares sense, the unregularized betas (shrinkage fraction equal to 1) obtained for that voxel.

To give a sense of the computational requirements of GLMsingle, we report here results for an example scenario. We ran the MATLAB version of GLMsingle with default parameters on the first NSD scan session for subj01 (1.8-mm standard-resolution version of the data). The scan session involved 750 trials and a data dimensionality of (81 voxels × 104 voxels × 83 voxels) = 699,192 voxels and (12 runs × 226 volumes) = 2,712 time points. The code was run on an 32-core Intel Xeon E5-2670 2.60 GHz Linux workstation with 128 GB of RAM and MATLAB 9.7 (R2019b). The data were loaded in single-precision format, resulting in a base memory usage of 8.4 GB of RAM (the data alone occupied 7.6 GB). Code execution (including figure generation and saving results to disk) took 4.8 hours (average of 2 trials). The maximum and mean memory usage over the course of code execution was 38.0 GB and 18.5 GB of RAM, respectively.

GLMsingle outputs

The default output from GLMsingle includes the different GLM beta estimates that are progressively obtained in the course of the algorithm (e.g. the single-trial betas with voxel-wise tailored HRFs; the single-trial betas incorporating GLMdenoise, etc.). The pipeline also outputs several metrics of interest, such as a map of the HRF indices chosen for different voxels, the R^2 values from the ON-OFF GLM, a map of the voxels identified as "noise", a summary plot of the cross-validation procedure used to select the number of noise regressors, and a map of the amount of ridge regression shrinkage applied at each voxel. These outputs are displayed in a set of convenient figures.

Flexibility of GLMsingle

Although GLMsingle provides default settings for the parameters that control its operation, the toolbox is flexible and allows the user to adjust the parameters if desired. Modifying the parameters allows the user to achieve a range of different behaviors, such as expanding the HRF library to include additional candidate HRFs; changing the maximum number of nuisance regressors tested during GLMdenoise (default is 10); modifying the range of shrinkage fractions evaluated for ridge regression (default is 0.05 to 1 in increments of 0.05); and running different flavors of GLM models that omit HRF fitting, GLMdenoise, and/or ridge regression. For complete documentation, please refer to the GLMsingle function descriptions and example scripts available at glmsingle.org.

Application of GLMsingle to NSD and BOLD5000

In order to assess the efficacy of GLMsingle for large-scale fMRI datasets, we tested GLMsingle on the NSD (Allen et al., 2022) and BOLD5000 (Chang et al., 2019) datasets. Both datasets involve presentation of many thousands of natural images. NSD and BOLD5000 share an overlapping subset of stimuli from the Microsoft Common Objects in Context (COCO) database (Lin et al., 2014), enabling direct comparison between the brain responses observed in the two datasets. However, there are a number of differences between the datasets: the two datasets were collected at different field strengths, with different event timings, and at different spatial and temporal resolution. In addition, while NSD contains many repeated stimuli within each scan session, BOLD5000 contains very few. As such, processing BOLD5000 requires grouping of input data across scan sessions to facilitate the cross-validation procedures used in GLMsingle. This challenging processing scheme with respect to image repetitions provides a strong test of the robustness of the GLMsingle technique.

NSD Dataset

For complete details of the NSD study, including scanning parameters, stimulus presentation, and experimental setup, refer to the *Methods* section of the corresponding dataset paper (Allen et al., 2022). In brief, a total of 8 subjects participated in the NSD experiment, each completing between 30-40 functional scanning sessions. For the full experiment, 10,000 distinct images from the Microsoft COCO dataset were designed to be presented 3 times each over the course of 40 sessions. For computational convenience and to make comparisons across subjects easier, only the first 10 NSD sessions from subjects 1–4 are used for the analyses contained in this manuscript. Functional data were collected at 7T, with 1.8-mm isotropic resolution, and with a TR of 1.6 s. Each trial lasted 4 s, and consisted of the presentation of an image for 3 s, followed by a 1-s gap. A total of 12 NSD runs were collected in one session, containing either 62 or 63 stimulus trials each, for a total of 750 trials per session.

The fMRI data from NSD were pre-processed by performing one temporal resampling to correct for slice time differences and one spatial resampling to correct for head motion within and across scan sessions, EPI distortion, and gradient nonlinearities. This procedure yielded volumetric fMRI time-series data in subject-native space for each NSD subject. In this paper, we analyze the standard-resolution pre-processed data from NSD which has 1.8-mm spatial resolution and 1.333-s temporal resolution (the time-series data are upsampled during preprocessing).

BOLD5000 Dataset

For complete details of the BOLD5000 study and methodology, refer to the corresponding dataset paper (Chang et al., 2019). A total of 4 subjects participated in the BOLD5000 dataset (CSI1-4). A full dataset contained 15 functional scanning sessions; subject CSI4 completed only 9 sessions before withdrawing from the study. BOLD5000 involved presentation of scene images from the Scene UNderstanding (SUN) (Xiao et al., 2010), COCO (Lin et al., 2014), and ImageNet (Deng et al., 2009) datasets. A total of 5,254 images, of which 4,916 images were unique, were used as the experimental stimuli. 112 of the 4,916 distinct images were shown four times and one image was shown three times to each subject. Functional data were collected at 3T, with 2-mm isotropic resolution, and with a TR of 2 s. Each trial lasted 10 s, and consisted of the presentation of an image for 1 s, followed by a 9-s gap. A total of either 9 or 10 runs were collected in one session, containing 37 stimulus trials each, for a total of either 333 or 370 trials per session.

The fMRI data from BOLD5000 were preprocessed using fMRIPrep (Esteban et al., 2019). Data preprocessing included motion correction, distortion correction, and co-registration to anatomy (or further details, please refer to the BOLD5000 dataset paper (Chang et al., 2019). This yielded volumetric fMRI time-series data in subject-native space for each BOLD5000 subject.

Because GLMsingle requires condition repetitions in order to perform internal cross-validation procedures, and because BOLD5000 contains a limited number of within-session repetitions, we concatenated data from groups of 5 sessions together before processing using GLMsingle. To account for differences in BOLD signal intensity across different sessions, we performed a global rescaling operation to the data within each session to roughly equate the time-series mean and variance across the 5 sessions comprising one batch of data. Specifically, we first computed the global mean fMRI volume across all 5 sessions, and then, for each session, computed a linear fit between the mean volume from a single session and the global mean volume. This yielded a multiplicative scaling factor applied to each session in order to roughly equate signal intensities across sessions.

Applying GLMsingle to NSD and BOLD5000

We used GLMsingle to estimate single-trial BOLD responses in the NSD and BOLD5000 datasets. For NSD, GLMsingle was applied independently to each scan session. For BOLD5000, groups of 5 sessions were processed together, following the rescaling procedure described above. The default GLMsingle parameters were used for processing both NSD and BOLD5000, except that we evaluated up to 12 nuisance regressors in GLMdenoise (default: 10).

Four different versions of single-trial GLM betas were computed and saved. The first beta version ($b1$, **AssumeHRF**) is the result of Step 2 of the GLMsingle algorithm, and reflects the use of a canonical HRF with no extra optimizations. We treat these generic GLM outputs as a baseline against which beta versions are compared; estimating BOLD responses using a canonical HRF and ordinary least squares (OLS) regression reflects an approach that has been commonly applied in the field of human neuroimaging. The second beta version ($b2$, **FitHRF**) is the result of Step 3, and reflects the result of voxel-wise HRF estimation. The third beta version ($b3$, **FitHRF + GLMdenoise**) is the result of Step 4, incorporating GLMdenoise, and the final beta version ($b4$, **FitHRF + GLMdenoise + RR**) arises from

Step 5, and reflects the additional use of ridge regression. For comparisons between GLMsingle and Least-Squares Separate (LSS) signal estimation (**Figure 3**), 4 auxiliary beta versions were computed. LSS betas were compared to those estimated using fractional ridge regression in the scenario of using the canonical HRF (**AssumeHRF + LSS** vs. **AssumeHRF + RR**) and in the scenario of performing HRF optimization using the GLMsingle library (**FitHRF + LSS** vs. **FitHRF + RR**). Our validation analyses involve comparing optimized GLMsingle betas ($b2$, $b3$, $b4$) against those estimated using the baseline GLM approach ($b1$), and performing an 8-way comparison incorporating both $b1$-$b4$ and the 4 auxiliary beta versions used for comparisons with LSS. Prior to all analyses, the responses of each voxel were z-scored within each experimental session in order to eliminate potential nonstationarities arising over time, and to equalize units across voxels.

Assessing the impact of GLMsingle

Analysis of voxel reliability

Computing test-retest reliability – To compute reliability, we repeated the following procedure for each beta version. We first extracted the betas from trials that correspond to repetitions of the same stimuli (NSD: 3 instances per stimulus; BOLD5000: 4 instances for subjects CSI1-3, and 3 for CSI4). For each voxel, this yielded a matrix of dimensions (repetitions x images). To compute reliability, Pearson correlation was computed between the average voxel response profiles for each possible unique split-half of the data. Therefore, in the case of 4 available repetitions, the reliability for a voxel was the average of 3 correlation values, with image repetitions grouped as follows: $corr(mean(1,2)$ vs. $mean(3,4))$; $corr(mean(1,3)$ vs. $mean(2,4))$; $corr(mean(1,4)$ vs. $mean(2,3))$. In the case of 3 repetitions, the reliability was the average of: $corr(mean(1,2)$ vs. $(3))$; $corr(mean(1,3)$ vs. $(2))$; $corr(mean(2,3)$ vs. $(1))$.

ROI analysis within visual cortex – To summarize reliability outcomes for each beta version, we used a liberal mask containing voxels in visual cortex. Specifically, we used the 'nsdgeneral' ROI from the NSD study, which was manually drawn on fsaverage to cover voxels responsive to the NSD experiment in the posterior aspect of cortex (Allen et al., 2022). To achieve a common reference ROI in volumetric space for each subject, we first transformed the nsdgeneral ROI to MNI space, and then mapped this ROI from MNI space to the space of each subject in NSD and each subject in BOLD5000.

Composite voxel reliability scores – In comparing different beta versions output by GLMsingle, we sought to understand whether the optimizations tended to affect all voxels equally, or whether the impact was mediated by voxel reliability. We therefore measured how different beta versions differed in our key outcome metrics (e.g. mean voxel reliability) as a function of the reliability of included voxels. To achieve fair comparisons, we ensured that the same groups of voxels were compared at each reliability threshold across beta versions. We achieved this by computing composite voxel reliability scores: the mean reliability value in each voxel over beta versions $b1$-$b4$. We then subselected groups of voxels by applying varying threshold levels to the composite reliability scores. For analyses incorporating the 4 auxiliary beta versions, composite reliability scores were computed as the mean across all 8 beta versions.

Effect of reliability on beta quality – To quantify the performance of different beta versions as a function of voxel reliability, composite scores were thresholded at increasing values (from Pearson $r = -0.2$ to 0.6, in steps of 0.05) to determine the included voxels at each reliability level. At each threshold, we computed the difference between the reliability achieved by a given beta version and the composite reliability (i.e. the average across beta versions). This difference was averaged across voxels, producing traces that reflect the relative quality of data from each beta version compared to the group average, across different levels of voxel reliability (**Figure 2b**).

Out-of-sample reliability analysis – GLMsingle makes use of all of the data that it is presented with, via a series of internal cross-validation operations. As such, there is some degree of dependence between runs. Note that this does not pose a significant "circularity" problem with respect to downstream analyses, as GLMsingle has no access to any scientific hypotheses and it is unlikely that GLMsingle could bias the single-trial beta estimates in favor of one hypothesis over another. However, when the primary analysis outcome is to establish that responses to the same condition are reliable across trials (e.g. **Figures 2, 3**), then that outcome is exactly what the GLMsingle algorithm is trying to achieve during hyperparameter selection. For a stringent quantification of reliability, we performed additional analyses in which quantification of reliability is restricted to responses estimated in completely independent calls to GLMsingle (**Figure 3b**). Specifically, we identify all instances where a condition is repeated within the same partition of data processed by GLMsingle (partition size: 1 session for NSD, 5 sessions for BOLD5000), and remove these instances from the calculation of reliability. The results show that even with strict separation, the patterns of results are essentially the same.

Comparison to LSS - Least-Squares Separate (LSS) is a popular technique for robust signal estimation in rapid event-related designs (Mumford et al., 2012, 2014; Abdulrahman and Henson, 2016). The LSS procedure fits a separate GLM for each stimulus, where the trial of interest is modeled as one regressor, and all other (non-target) trials are collapsed into a second regressor. An implementation of LSS is included in the GLMsingle toolbox.

Analysis of temporal autocorrelation

A commonly used strategy to increase fMRI statistical power is to increase the number of experimental trials by allowing them to be presented close together in time. However, given the sluggish nature of BOLD responses and the existence of temporal noise correlations, this strategy tends to yield correlations in GLM beta estimates for nearby trials (Mumford et al., 2014; Olszowy et al., 2019; Woolrich et al., 2001; Kumar and Feng, 2014). In general, we expect that such correlations are largely artifactual and unwanted. Given that GLMsingle attempts to reduce noise levels, we sought to explore whether GLMsingle has a noticeable impact on temporal autocorrelation.

Average temporal autocorrelation by dataset – For each beta version, the following procedure was used to assess the degree of temporal autocorrelation in the data. For visual cortex data from each experimental session (nsdgeneral ROI, Allen et al., 2022), we computed the Pearson correlation between the spatial response patterns from each pair of trials in the session, yielding a representational similarity matrix (RSM) where the temporal ordering of trials is preserved. This process was repeated for all sessions, yielding a total of 10 RSMs for each NSD subject and 15 RSMs for each BOLD5000 subject (9 for subject CSI4, who did not complete the full study). To assess autocorrelation in the data – relationships arising due to temporal proximity of different trials – we then took the average of all RSMs within each dataset. Note that in both NSD and BOLD5000, the order of stimulus presentation was essentially unstructured (pseudorandom). Thus, in terms of signal content (stimulus-driven responses in the absence of noise), we expect that trials should be uncorrelated, on average, and that any non-zero correlations are indicative of the effects of noise that persist following GLM fitting. The extent to which non-zero r values extend forward in time from the RSM diagonal indicates the timescale of the noise effects in a given beta version.

Computing the autocorrelation function – For quantitative summary, we computed a temporal autocorrelation function from the dataset-averaged RSM for each beta version (**Figure 4**). For a given RSM, we computed the average similarity value between all trials k and $k + x$, where x varies from 1 to

n, where n is the dimensionality of the RSM. Intuitively, at $x = 1$, $autocorr(x)$ equals the average of all values falling 1 index below the diagonal of the RSM; at $x = 5$, it equals the average of all values falling 5 indices below the diagonal, etc. This procedure outputs a succinct summary of the average correlation in neural response between all pairs of time-points within a session, allowing for easy comparison between the beta versions in a single plot (**Figure 4**, right-most column). The theoretical desired outcome is $autocorr(x) = 0$; thus, beta versions whose autocorrelation functions are "flatter" (e.g. less area under the curve) presumably contain more accurate GLM estimates. Because the temporal interval between trials differed between NSD (4 s) and BOLD5000 (10 s), we express the autocorrelation functions in terms of seconds post-stimulus for plotting, to allow for straightforward comparison between the datasets.

Effect of reliability on temporal autocorrelation – The effect of temporal autocorrelation in GLM betas may vary depending on the relative responsiveness of different voxels to the experimental stimuli. As such, we repeated the autocorrelation analyses several times, varying the expanse of voxels that were included. We again relied on the aggregate reliability scores (computed previously) as a measure of voxel quality, which are the average voxel reliabilities taken across all the beta versions under consideration. This avoids biasing the voxel selection procedure. In **Figure 4**, we compare temporal autocorrelation trends arising from analysis of voxels at two different reliability thresholds ($r = 0$ and $r = 0.3$).

Analysis of between-subject representational similarity

Another way to assess the quality of beta estimates is to examine the similarity of BOLD response estimates across subjects. The underlying logic is that noise is expected to be stochastic in the data acquisition for each subject, and thus, should on average increase the dissimilarities of BOLD response estimates across subjects. A method that accurately removes noise would then be expected to increase the similarity of BOLD responses across subjects. To quantify response similarity, we use representational similarity analysis (RSA), a commonly used approach in systems and cognitive neuroscience (Kriegeskorte et al., 2008; Nili et al., 2014; Diedrichsen and Kriegeskorte, 2017; Kaniuth and Hebart, 2021).

Between-subject RSA correlations – For comparisons between subjects across NSD and BOLD5000, we identified a subset of 241 images that overlapped between BOLD5000 and the portion of NSD being analyzed for this manuscript. Once overlapping images were identified, the corresponding GLM betas for each version were isolated, and averaged over all available repetitions within subject (3 for NSD, 4 for BOLD5000). Then, we used Pearson dissimilarity ($1 - r$) to compute RDMs over the averaged betas for each subject, in each dataset. To assess the impact of voxel reliability on cross-subject RDM correlations, this procedure was repeated across a range of voxel reliability inclusion levels $r = [-1, 0, 0.05, 0.1, 0.15, 0.2, 0.25]$, using the beta version-averaged aggregate reliability scores computed previously. Voxels inside the nsdgeneral ROI were used in this analysis. Once RDMs were computed for each subject, using responses from the sets of stimuli detailed above, within- and across-dataset RSA correlations were computed using the vectorized lower-triangular portions of each RDM (**Figure 5b**).

Analysis of MVPA decoding accuracy

Multivoxel pattern analysis (MVPA) investigates the information contained in distributed patterns of neural activity to infer the functional role of brain areas and networks. Pattern decoding tools like MVPA have been deployed extensively in systems and cognitive neuroscience to study the function of neural ROIs (Haxby et al., 2001; Norman et al., 2006; Naselaris et al., 2011; Charest et al., 2018). To further assess the practical impact of GLMsingle, we tested the efficacy of MVPA decoding using the different beta versions output by the toolbox.

Image-level decoding paradigm – We implemented a challenging "one-vs-many" decoding task to assess whether data quality was sufficiently high to characterize the distinct neural patterns associated with individual naturalistic images in the NSD and BOLD5000 datasets. Within each dataset, we identified the set of images that all subjects viewed at least 3 times, and then performed multiclass linear support vector machine (SVM) decoding via leave-one-repetition-out cross-validation. In NSD, a total of 82 classes were used, representing the images that overlapped across the 10 available sessions from subj01-04. In BOLD5000, the subset of these 82 stimuli overlapping between all subjects of both datasets were used (a total of 20 classes). We then assessed the degree to which relative differences in decoding accuracy between $b1$ and $b4$ changed depending on the reliability of the included voxels. We conducted the above decoding procedure iteratively, each time increasing the voxel reliability inclusion threshold for data within the nsdgeneral ROI (range r = 0 to 0.35). BOLD5000 subject CSI4, having completed only 9 of 15 experimental sessions, was excluded from MVPA procedures due to insufficient stimulus repetitions.

Multidimensional scaling – To gain insight into the representational changes due to GLMsingle that may support improvements in MVPA decoding, we performed multidimensional scaling (MDS) over repetition-averaged NSD betas from a baseline GLM ($b1$) and the final betas from GLMsingle ($b4$), within the nsdgeneral ROI of an example subject (NSD subj01). In **Figure 6b**, we compare the 2-dimensional MDS embeddings between these beta versions, coloring COCO stimuli based on whether they contain animate or inanimate objects according to the image annotations.

Acknowledgments

Collection of the NSD dataset was supported by NSF CRCNS grants IIS-1822683 (to K.K.) and IIS-1822929 (to Thomas Naselaris). We thank N. Blauch, A. Wang, E. Aminoff, and R. River for helpful discussions.

Author Contributions

KNK, JAP, and MJT led the fMRI studies yielding data analyzed here. JSP devised and performed the analyses. IC and KNK implemented the GLMsingle technique in Python and MATLAB, respectively. JSP and JWK created the GLMsingle online example scripts. JSP and KNK wrote the manuscript. All authors discussed the results and provided feedback on the manuscript.

Conflict of Interest Statement

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

References

Abdulrahman, H. and Henson, R. N. (2016). Effect of trial-to-trial variability on optimal event-related fmri design: Implications for beta-series correlation and multi-voxel pattern analysis. *NeuroImage*, 125:756–766.

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. (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126.

Bai, B. and Kantor, P. (2007). A shape-based finite impulse response model for functional brain images. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 440–443. IEEE.

Bao, P., She, L., McGill, M., and Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108.

Blauch, N. M., Behrmann, M., and Plaut, D. C. (2021). A connectivity-constrained computational account of topographic organization in high-level visual cortex. *bioRxiv*.

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., and Aminoff, E. M. (2019). Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):1–18.

Charest, I., Kriegeskorte, N., and Kay, K. N. (2018). Glmdenoise improves multivariate pattern analysis of fmri data. *NeuroImage*, 183:606–616.

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Hum Brain Mapp*, 8(2-3):109–114.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

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

Doshi, F. and Konkle, T. (2021). Organizational motifs of cortical responses to objects emerge in topographic projections of deep neural networks. *Journal of Vision*, 21(9):2226–2226.

D'Isanto, A. and Polsterer, K. L. (2018). Photometric redshift estimation via deep learning-generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy & Astrophysics*, 609:A111.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019). fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, 16(7):974–981.

Handwerker, D. A., Gonzalez-Castillo, J., D'esposito, M., and Bandettini, P. A. (2012). The continuing challenge of understanding and modeling hemodynamic variation in fmri. *Neuroimage*, 62(2):1017–1023.

Handwerker, D. A., Ollinger, J. M., and D'Esposito, M. (2004). Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, 21(4):1639–1651.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., and Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Horikawa, T. and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nat Commun*, 8:15037.

Iten, R., Metger, T., Wilming, H., Del Rio, L., and Renner, R. (2020). Discovering physical concepts with neural networks. *Physical review letters*, 124(1):010508.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Kaniuth, P. and Hebart, M. N. (2021). Feature-reweighted rsa: A method for improving the fit between computational models, brains, and behavior. *bioRxiv*.

Kay, K. (2022). The risk of bias in denoising methods. *arXiv preprint arXiv:2201.09351*.

Kay, K., Jamison, K. W., Zhang, R.-Y., and Uğurbil, K. (2020). A temporal decomposition method for identifying venous effects in task-based fmri. *Nature methods*, 17(10):1033–1039.

Kay, K., Rokem, A., Winawer, J., Dougherty, R., and Wandell, B. (2013). Glmdenoise: a fast, automated technique for denoising task-based fmri data. *Frontiers in neuroscience*, 7:247.

Kay, K. N., David, S. V., Prenger, R. J., Hansen, K. A., and Gallant, J. L. (2008). Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fmri. Technical report, Wiley Online Library.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Kriegeskorte, N. and Wei, X.-X. (2021). Neural tuning and representational geometry. *arXiv preprint arXiv:2104.09743*.

Kumar, A. and Feng, L. (2014). Efficient regularization of temporal autocorrelation estimates in fmri data. In *The 15th International Conference on Biomedical Engineering*, pages 88–91. Springer.

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L., and DiCarlo, J. J. (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*.

Li, S. P. D. and Bonner, M. F. (2021). Tuning in scene-preferring cortex for mid-level visual features gives rise to selectivity across multiple levels of stimulus complexity. *bioRxiv*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Liu, T. T. (2016). Noise contributions to the fmri signal: An overview. *NeuroImage*, 143:141–151.

Liu, T. T., Frank, L. R., Wong, E. C., and Buxton, R. B. (2001). Detection power, estimation efficiency, and predictability in event-related fmri. *Neuroimage*, 13(4):759–773.

Long, B., Yu, C.-P., and Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38):E9015–E9024.

Marques, T., Schrimpf, M., and DiCarlo, J. J. (2021). Multi-scale hierarchical neural network models that bridge from single neurons in the primate primary visual cortex to object recognition behavior. *bioRxiv*.

Monti, M. M. (2011). Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. *Front Hum Neurosci*, 5:28.

Mumford, J. A., Davis, T., and Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *Neuroimage*, 103:130–138.

Mumford, J. A., Turner, B. O., Ashby, F. G., and Poldrack, R. A. (2012). Deconvolving bold activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, 59(3):2636–2643.

Naselaris, T., Allen, E., and Kay, K. (2021). Extensive sampling for complete models of individual brains. *Current Opinion in Behavioral Sciences*, 40:45–51.

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

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553.

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430.

Olszowy, W., Aston, J., Rua, C., and Williams, G. B. (2019). Accurate autocorrelation modeling substantially improves fmri reliability. *Nature communications*, 10(1):1–11.

Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., and Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11):1628–1636.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al. (2021). Skillful precipitation nowcasting using deep generative models of radar. *arXiv preprint arXiv:2104.00954*.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770.

Rokem, A. and Kay, K. (2020). Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. *GigaScience*, 9(12):giaa133.

Schawinski, K., Turp, M. D., and Zhang, C. (2018). Exploring galaxy evolution with generative models. *Astronomy & Astrophysics*, 616:L16.

Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, 5:399–426.

Soch, J., Allefeld, C., and Haynes, J.-D. (2020). Inverse transformed encoding models–a solution to the problem of correlated trial-by-trial parameter estimates in fmri decoding. *Neuroimage*, 209:116449.

Wilson, M. (2018). What is smart hdr? explaining apple's new camera tech — trusted reviews. https://www.trustedreviews.com/news/what-is-smart-hdr-3565603. (Accessed on 12/22/2021).

Woolrich, M. W., Ripley, B. D., Brady, M., and Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fmri data. *Neuroimage*, 14(6):1370–1386.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624.

Zhang, Y., Zhou, K., Bao, P., and Liu, J. (2021). Principles governing the topological organization of object selectivities in ventral temporal cortex. *bioRxiv*.