

¹ GLMsingle: a toolbox for improving single-trial fMRI response estimates

² 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

³ ¹*Department of Psychology, Harvard University, Cambridge, MA, USA*

⁴ ²*Center for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, UK*

⁵ ³*cerebrUM, Département de Psychologie, Université de Montréal, Montréal, Canada*

⁶ ⁴*Department of Psychology, New York University, New York, NY, USA*

⁷ ⁵*Center for Human Neuroscience, Department of Psychology, University of Washington, Seattle, WA, USA*

⁸ ⁶*Department of Psychology, Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA*

⁹ ⁷*Center for Magnetic Resonance Research (CMRR), Department of Radiology, University of Minnesota, Minneapolis, MN, USA*

¹¹ * Corresponding author (jacob.samuel.prince@gmail.com)

¹² ABSTRACT

¹³ **Advances in modern artificial intelligence (AI) have inspired a paradigm shift in human neuroscience,**
¹⁴ **yielding large-scale functional magnetic resonance imaging (fMRI) datasets that provide high-resolution**
¹⁵ **brain responses to tens of thousands of naturalistic visual stimuli. Because such experiments necessarily**
¹⁶ **involve brief stimulus durations and few repetitions of each stimulus, achieving sufficient signal-to-noise**
¹⁷ **ratio can be a major challenge. We address this challenge by introducing *GLMsingle*, a scalable,**
¹⁸ **user-friendly toolbox available in MATLAB and Python that enables accurate estimation of single-trial**
¹⁹ **fMRI responses (glmsingle.org). Requiring only fMRI time-series data and a design matrix as inputs,**
²⁰ ***GLMsingle* integrates three techniques for improving the accuracy of trial-wise general linear model**
²¹ **(GLM) beta estimates. First, for each voxel, a custom hemodynamic response function (HRF) is identified**
²² **from a library of candidate functions. Second, cross-validation is used to derive a set of noise regressors**
²³ **from voxels unrelated to the experimental paradigm. Third, to improve the stability of beta estimates for**
²⁴ **closely spaced trials, betas are regularized on a voxel-wise basis using ridge regression. Applying**
²⁵ ***GLMsingle* to the Natural Scenes Dataset and BOLD5000, we find that *GLMsingle* substantially improves**
²⁶ **the reliability of beta estimates across visually-responsive cortex in all subjects. Furthermore, these**
²⁷ **improvements translate into tangible benefits for higher-level analyses relevant to systems and cognitive**
²⁸ **neuroscience. Specifically, we demonstrate that *GLMsingle*: (i) improves the decorrelation of response**
²⁹ **estimates between trials that are nearby in time; (ii) enhances representational similarity between subjects**
³⁰ **both within and across datasets; and (iii) boosts one-versus-many decoding of visual stimuli. *GLMsingle* is**
³¹ **a publicly available tool that can significantly improve the quality of past, present, and future**
³² **neuroimaging datasets that sample brain activity across many experimental conditions.**

³³ **Keywords:** fMRI preprocessing, GLM, large-scale datasets, denoising, voxel reliability

³⁴ INTRODUCTION

³⁵ Across many scientific disciplines, datasets are rapidly increasing in size and scope. These resources
³⁶ have kickstarted a new era of data-driven scientific discovery (Richards et al., 2019; Jumper et al.,
³⁷ 2021; Iten et al., 2020; Ravuri et al., 2021; Schawinski et al., 2018; D’Isanto and Polsterer, 2018).
³⁸ In visual neuroscience, recent efforts to sample individual brains at unprecedented scale and depth
³⁹ have yielded high-resolution functional magnetic resonance imaging (fMRI) datasets in which subjects
⁴⁰ view thousands of distinct images over several dozen hours of scanning (see Naselaris et al., 2021 for
⁴¹ a review). These exciting “condition-rich” datasets are large enough to propel the development of
⁴² computational models of how humans process complex naturalistic stimuli. For example, resources
⁴³ such as the Natural Scenes Dataset (NSD, Allen et al., 2022), BOLD5000 (Chang et al., 2019), and
⁴⁴ THINGS (Hebart et al., 2019) may be useful for advancing our ability to characterize the tuning (Bao
⁴⁵ et al., 2020; Li and Bonner, 2021; Long et al., 2018; Kriegeskorte and Wei, 2021; Popham et al., 2021),

46 topography (Blauch et al., 2021; Doshi and Konkle, 2021; Zhang et al., 2021; Lee et al., 2020), and
47 computations (Yamins et al., 2014; DiCarlo et al., 2012; Freeman et al., 2013; Marques et al., 2021;
48 Horikawa and Kamitani, 2017) performed in visual cortex.

49 The potential of large-scale datasets to reveal general principles of neural function depends critically on
50 signal-to-noise ratio (SNR), which refers to one's ability to reliably measure distinct neural signatures
51 associated with different stimuli or experimental conditions. Diverse sources of noise affect fMRI data,
52 and these noise sources limit the robustness and interpretability of data analyses (Liu, 2016; Kay et al.,
53 2013). For example, subject head motion, scanner instabilities, physiological noise, and thermal noise
54 all contribute unwanted variability to fMRI data. Noise is especially problematic in studies that sample
55 a large number of conditions, since the number of repetitions of each condition is typically limited,
56 resulting in noisy responses even after trial-averaging.

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

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

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

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

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

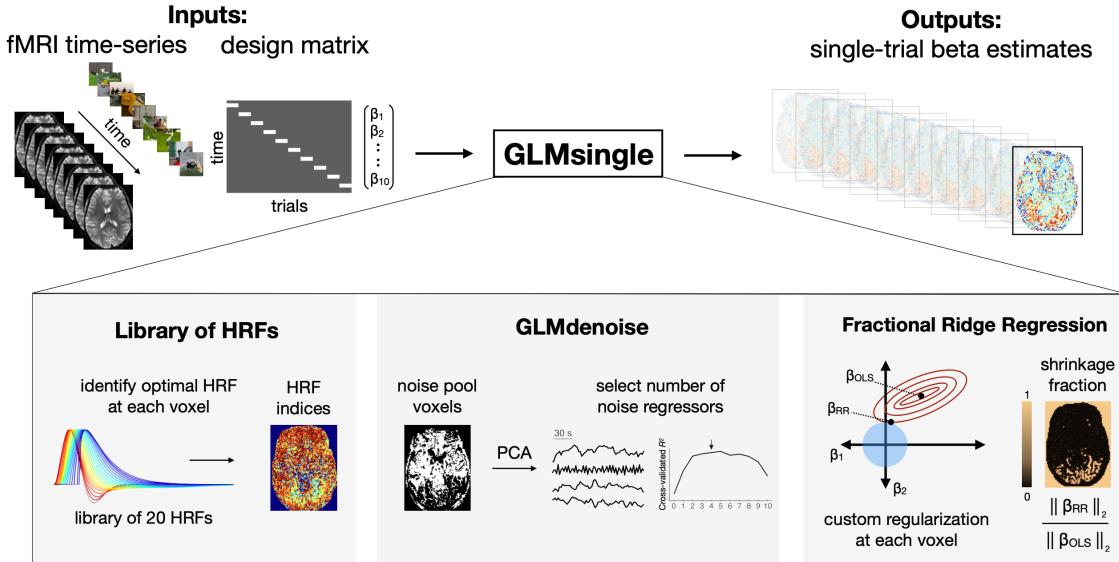


Figure 1: Overview of GLMsingle

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.

94 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:

- 101 1. A baseline single-trial GLM is used to model each stimulus trial separately using a canonical
102 HRF. This provides a useful baseline for comparison (*b1: AssumeHRF*).
- 103 2. An optimal HRF is identified for each voxel (Allen et al., 2022) by iteratively fitting a set
104 of GLMs, each time using a different HRF from a library of 20 HRFs. For each voxel, we

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

107 3. GLMdenoise (Kay et al., 2013; Charest et al., 2018) is used to determine nuisance regressors to
 108 include in the model. Principal components analysis is applied to time-series data from a pool of
 109 noise voxels (see *Methods* for details), and the top principal components are added one at a time
 110 to the GLM until cross-validated variance explained is maximized on-average across voxels (*b3: FitHRF + GLMdenoise*).
 111

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

115 **GLMsingle improves the reliability of beta estimates**

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

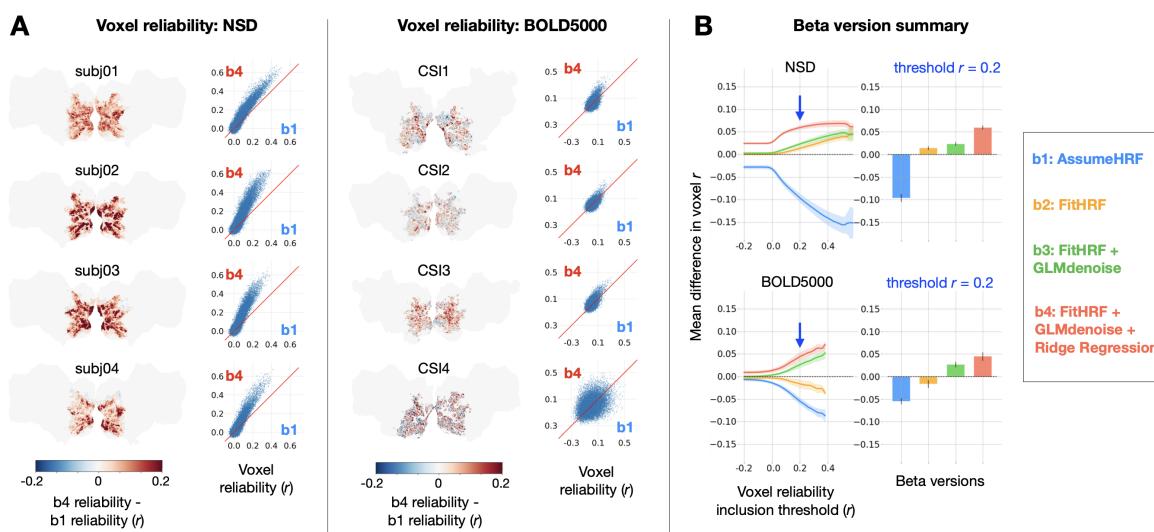


Figure 2: Impact of GLMsingle on voxel test-retest reliability

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 (\pm 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.

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

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

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

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

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

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

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

160 **GLMsingle helps disentangle neural responses to neighboring trials**

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

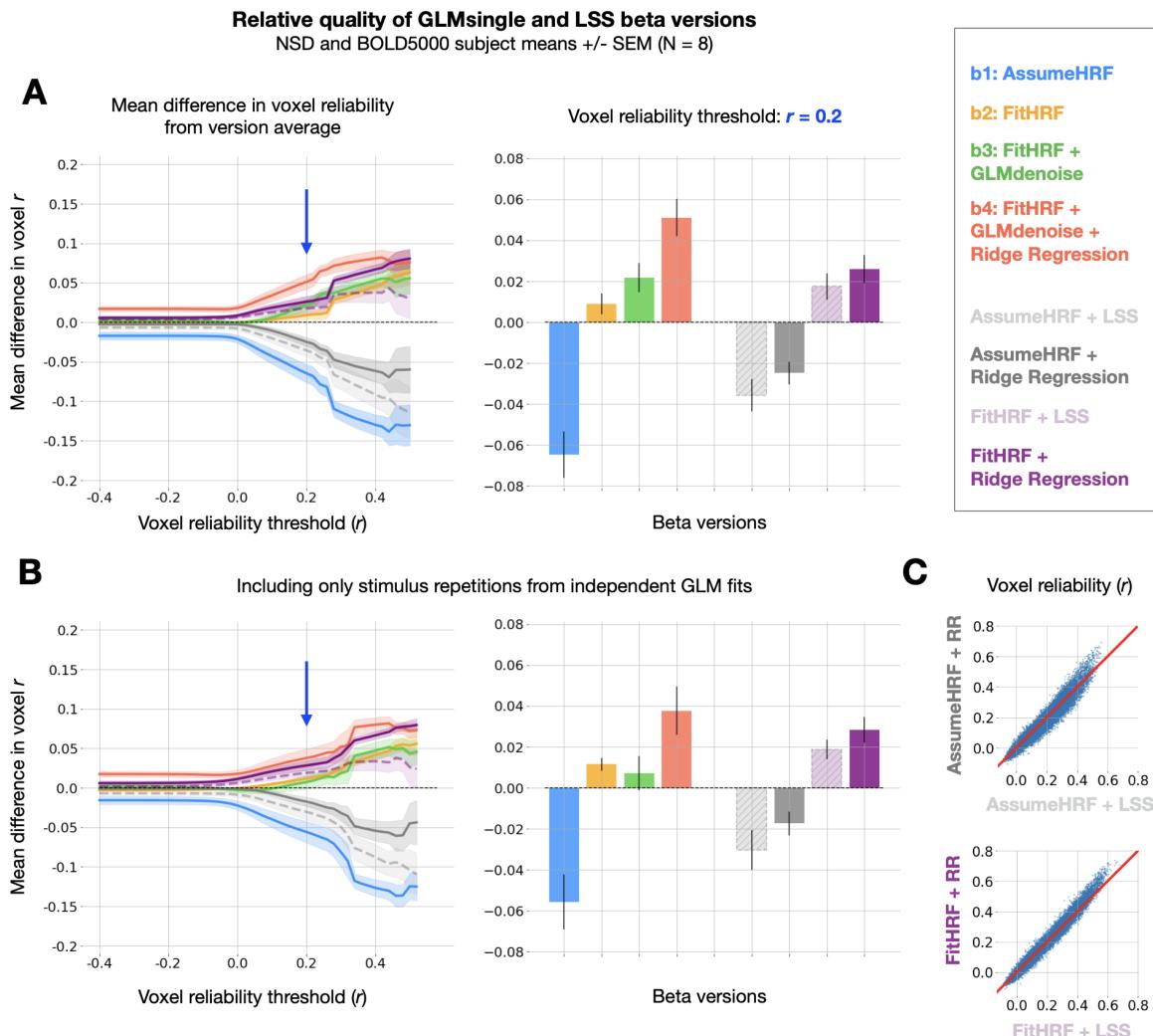


Figure 3: Comparison between GLMsingle and LSS

(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.

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

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

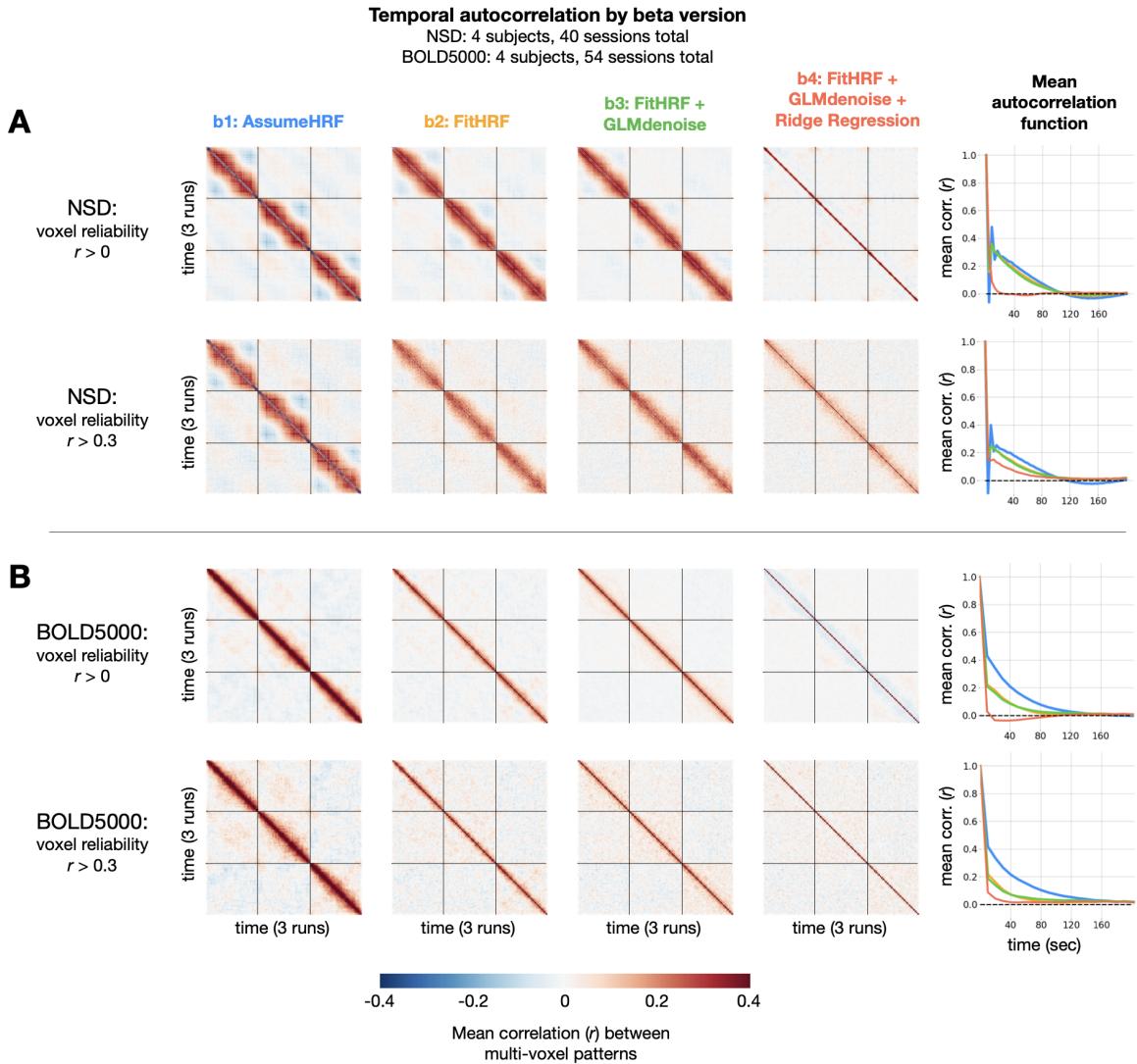


Figure 4: Impact of GLMsingle on temporal autocorrelation

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*

182 for $b1$, a known artifact of OLS fitting in the case of high multicollinearity between GLM predictors
183 (Mumford et al., 2014; Soch et al., 2020).

184 When applying GLMsingle, these patterns of temporal autocorrelation change dramatically. In NSD
185 $b4$, autocorrelation drops to $r = 0$ much more rapidly than in $b1$, and in BOLD5000, beta maps from
186 successive trials in $b4$ are now nearly uncorrelated on average. This is an expected outcome, since
187 the stimuli in NSD and BOLD5000 are ordered pseudorandomly. In both datasets, an intermediate
188 beta version ($b2$) containing only HRF optimization confers marginal benefit over $b1$, but the most
189 dramatic improvements come from the addition of both GLMdenoise and fractional ridge regression
190 ($b4$). Overall, these results demonstrate the utility of GLMsingle for disentangling neural responses
191 to nearby stimuli in event-related designs, even when events are presented relatively slowly (as in
192 BOLD5000).

193 **GLMsingle improves between-subject representational similarity across datasets**

194 Large-scale datasets such as NSD and BOLD5000 are well-suited for representational analyses (e.g.,
195 RSA) that compare evoked neural response patterns between individual subjects, across different exper-
196 imental modalities, and against computational models (e.g., deep neural networks, see Kriegeskorte,
197 2015, Serre, 2019 for review.) In almost all such studies, representational analyses presume that the
198 same set of stimuli will evoke reasonably similar responses across subjects. As such, given the ubiquity
199 of noise in fMRI, it is reasonable to expect that improving the accuracy of single-trial response estimates
200 should yield representations that are more similar across individuals.

201 To compare representations between subjects, we used the approach of RSA (Kriegeskorte et al.,
202 2008). First, we isolated stimuli that overlap between BOLD5000 and the subset of NSD analyzed
203 for this manuscript (the first 10 sessions from each subject). Using these 241 stimuli, we constructed
204 representational dissimilarity matrices (RDMs) using repetition-averaged betas from each individual,
205 and then correlated all pairs of subject RDMs within and between datasets. Note that GLMsingle is not
206 designed to enhance or optimize cross-subject representational similarity; as such, it is informative to
207 examine RSA correlations between subjects as a way of assessing methods for denoising (Charest et al.,
208 2018). Strikingly, in comparing beta versions $b1$ and $b4$, we observe a consistent strengthening of RDM
209 correspondence (**Figure 5b**). This trend held within NSD, within BOLD5000, and when comparing the
210 RDMs of subject pairs between the two datasets. The latter result is especially striking given the many
211 methodological differences between NSD and BOLD5000: fMRI data were collected at different sites
212 on different scanners, at different field strengths (7T vs. 3T), with different behavioral tasks, and with
213 different inter-stimulus intervals (4 s vs. 10 s).

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

218 **GLMsingle enables fine-grained image-level MVPA decoding**

219 As a final analysis, we assessed the effect of GLMsingle on the results of multivoxel pattern analysis
220 (MVPA). In a “one-vs.-many” classification paradigm, we trained linear SVM models for each subject
221 to predict image identity from neural response patterns. The baseline GLM ($b1$) classification accuracy
222 was slightly above chance on average for the subjects in NSD and BOLD5000 when including all visual
223 cortex voxels (**Figure 6a**, blue traces). Performing the same MVPA procedure using GLMsingle betas
224 ($b4$), we observe that mean accuracy approximately triples in NSD and doubles in BOLD5000 (**Figure**
225 **6a**, red traces). Moreover, in both datasets we observe a substantial increase in classification accuracies

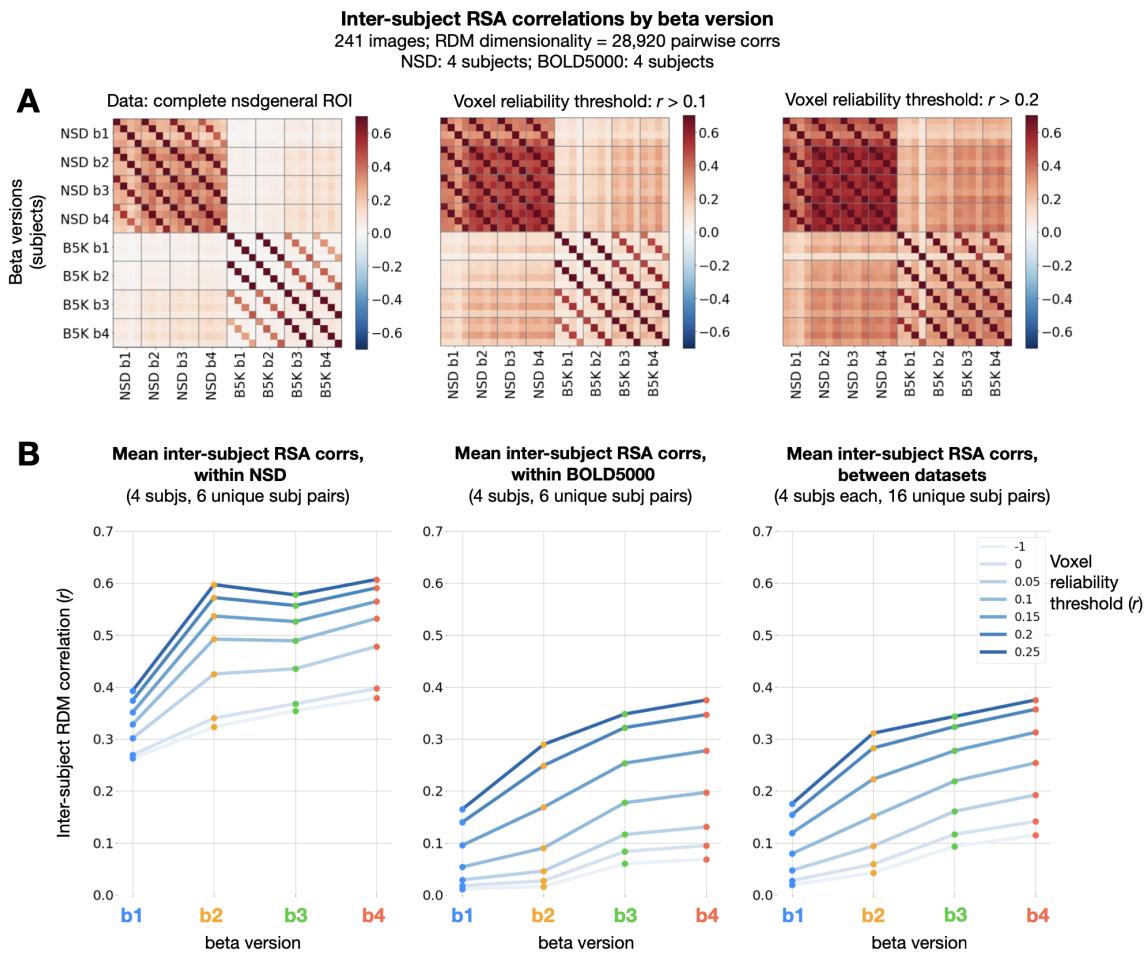


Figure 5: Impact of GLMsingle on inter-subject RSA correlations

(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.

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

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

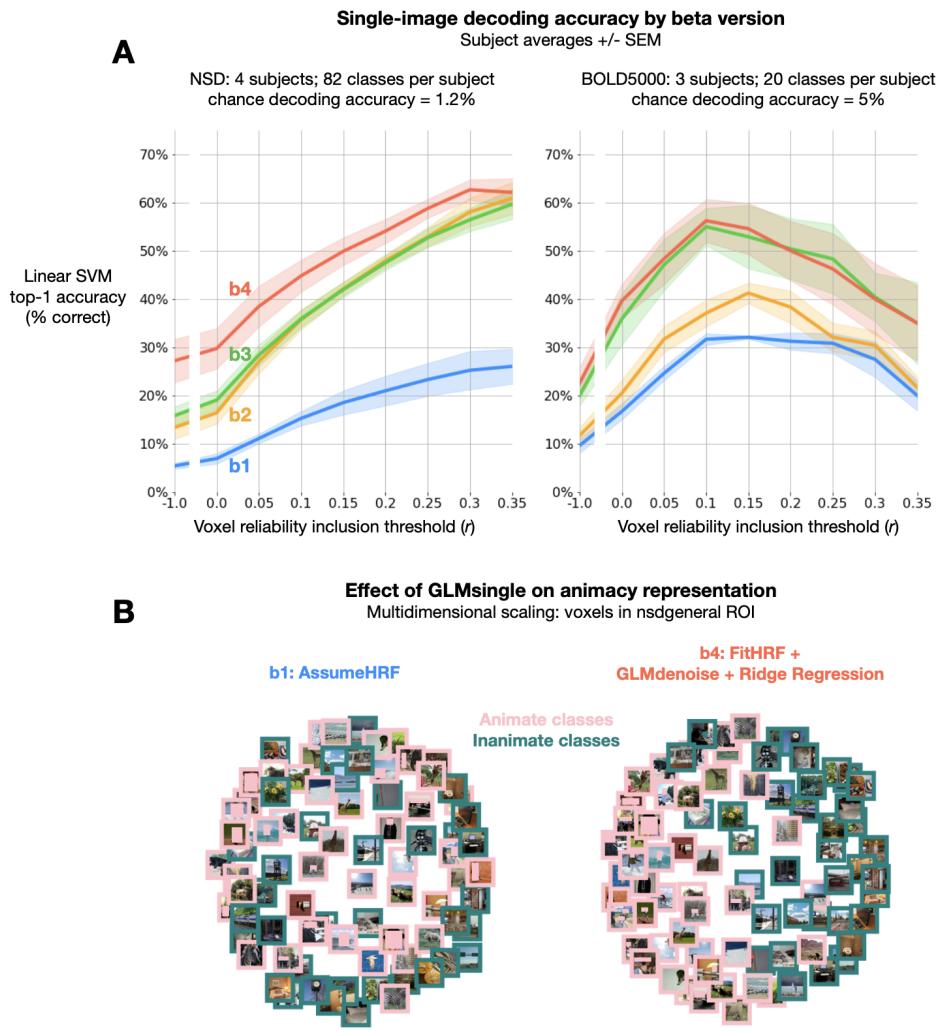


Figure 6: Impact of GLMsingle on MVPA decoding accuracy

(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.

234 **DISCUSSION**

235 As scientific datasets grow in scale and scope, new techniques for data processing will help to unlock
 236 their potential. This is especially true in human neuroscience where data remain both expensive and
 237 time-consuming to collect (Naselaris et al., 2021). This paper has introduced GLMsingle, a publicly
 238 available toolbox for analyzing fMRI time-series data that leverages data-driven techniques to improve
 239 the accuracy of single-trial fMRI response estimates. We have tested GLMsingle extensively using NSD
 240 and BOLD5000, two of the largest fMRI datasets that densely sample responses within individuals.
 241 For both datasets, analyses of the response estimates provided by GLMsingle indicate substantial
 242 improvements in several key metrics of interest to neuroscientists: (i) enhanced test-retest reliability of
 243 voxel response profiles, a straightforward metric of data quality; (ii) reduced temporal autocorrelation,

244 a common fMRI effect that is presumably undesirable and especially prominent in rapid event-related
245 designs; (iii) increased representational similarity across subjects both within and across datasets; and
246 (iv) improved multivariate pattern classification performance when discriminating responses evoked by
247 individual images.

248 **Principles underlying GLMsingle**

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

- 250 1. *HRF fitting.* GLMsingle uses a “library of HRFs” technique to select the most appropriate HRF
251 to use for each voxel in a given dataset (Allen et al., 2022). This library consists of a set of
252 20 HRFs that were derived from experimental data (specifically, the first NSD scan session
253 acquired in each of the 8 NSD subjects). It is well known that variations in HRFs exist across
254 voxels, brain areas, and subjects, and that mismodeling the timecourse of a voxel may lead to
255 suboptimal analysis outcomes (Handwerker et al., 2004, 2012). Imposing constraints on HRF
256 selection by choosing from a fixed set of HRFs avoids the instability (high variance) associated
257 with more flexible timecourse modeling approaches, such as finite impulse response modeling
258 (Kay et al., 2008; Bai and Kantor, 2007). Variations in timecourse shapes in the HRF library
259 reflect a continuum between short-delay, narrow-width timecourses to long-delay, broad-width
260 timecourses, and are likely caused by variations in the contribution of large vessels to the BOLD
261 response observed in a voxel (Kay et al., 2020).
- 262 2. *Data-driven denoising.* Incorporating an adaptation of the GLMdenoise technique (Kay et al.,
263 2013), GLMsingle uses principal components analysis to calculate potential nuisance regressors
264 from fMRI time-series data observed in voxels that are deemed unrelated to the experimental
265 paradigm. These regressors are incorporated into the GLM using a cross-validation procedure to
266 determine the optimal number of nuisance regressors to add. A key aspect of our approach is
267 the acknowledgement that including increasing numbers of nuisance regressors will, at some
268 point, cause overfitting and degradation of results (Kay et al., 2013); this motivates the use of
269 cross-validation to determine the optimal level of model complexity.
- 270 3. *Regularization of GLM weights.* To improve the accuracy of single-trial GLM response estimates,
271 GLMsingle uses fractional ridge regression (Rokem and Kay, 2020), with an optimal degree of
272 regularization identified for each voxel, again using cross-validation. The improvements afforded
273 by this procedure are due to the substantial amount of overlap of the fMRI response across
274 successive trials, unless very long (> 30 s) inter-stimulus intervals are used. It is well known
275 that, in the context of ordinary least squares estimation, two predictors that are correlated (or
276 anti-correlated) will have reduced estimation precision compared to the scenario in which the
277 predictors are uncorrelated (Mumford et al., 2012; Soch et al., 2020). For rapid event-related
278 designs, predictors for consecutive trials are typically correlated, and ordinary least-squares
279 estimates will suffer from high levels of instability. Ridge regression imposes a shrinkage prior
280 (penalizing the sum of the squares of the beta estimates), which can, in principle, dampen the
281 effects of noise and improve out-of-sample generalizability of the beta estimates.

282 **Ideal use-cases for GLMsingle**

283 GLMsingle is designed to be general in its application. It uses data-driven procedures that automatically
284 adapt to the signal-to-noise characteristics of a given dataset. For example, in datasets where structured
285 noise is prevalent, appropriate nuisance regressors will automatically be included, whereas in datasets
286 with very little structured noise (e.g., low head motion), fewer (or no) nuisance regressors will be

287 included. As another example, for experimental designs with high temporal overlap between consecutive
288 trials or high levels of noise, relatively strong levels of shrinkage regularization will likely be selected.

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

293 **Potential limitations to consider when applying GLMsingle**

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

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

313 It is important to consider whether denoising comes at the potential cost of introducing bias (Kay,
314 2022). Considering each component of GLMsingle, we believe that the risk of bias is minimal for most
315 use cases. First, considering the library-of-HRFs approach, we note that the conventional approach
316 of using a fixed canonical HRF actually incurs more risk of biasing response estimates than does an
317 approach that attempts to flexibly capture variations in HRFs. Nonetheless, we acknowledge that the
318 library may not necessarily capture all HRF shapes, and this represents one possible source of bias
319 (though likely minor). Second, considering the GLMdenoise procedure, we note that data-derived
320 nuisance regressors are not blindly removed from the time-series data prior to modeling, as this would
321 pose a clear risk of removing experimentally-driven signals, thereby leading to bias (Liu et al., 2001).
322 Rather, by including both task-related regressors and nuisance regressors in the GLM, the model can
323 appropriately partition variance between signal and noise sources. Third, considering ridge regression,
324 we note that shrinkage can be viewed as a form of temporal smoothing, in the sense that beta weights
325 from temporally adjacent trials are biased to be more similar in magnitude. While this is indeed a
326 source of bias, this should be concerning only for investigations where relative responses for nearby
327 trials are of specific interest (e.g., studies of repetition suppression). For other investigations, and
328 especially for experiments where condition ordering is pseudorandom, it is unlikely that this form of
329 temporal regularization and its associated bias would lead to incorrect scientific inferences.

330 **Online example scripts and tutorials**

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

339 **Conclusion**

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

351 **MATERIALS AND METHODS**

352 **Description of GLMsingle**

353

354 **Inputs to GLMsingle**

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

366 **GLMsingle overview**

367 GLMsingle consists of three main analysis components. The first component is the use of a library of
368 hemodynamic response functions (HRFs) to identify the best-fitting HRF for each voxel. This simple
369 approach for compensating for differences in hemodynamic timecourses across voxels ([Handwerker
370 et al., 2004](#)) has several appealing features: it invariably provides well-regularized HRF estimates, and
371 it is efficient and can be executed with reasonable computational cost. The second component is an
372 adaptation of GLMdenoise to a single-trial GLM framework. GLMdenoise is a previously introduced
373 technique ([Kay et al., 2013](#)) in which data-derived nuisance regressors are identified and used to remove

374 noise from—and therefore improve the accuracy of—beta estimates. The third analysis component is an
375 application of ridge regression (Hoerl and Kennard, 1970) as a method for dampening the noise inflation
376 caused by correlated single-trial GLM predictors. To determine the optimal level of regularization for
377 each voxel, we make use of a recently developed efficient re-parameterization of ridge regression called
378 “fractional ridge regression” (Rokem and Kay, 2020).

379 **Derivation of the library of HRFs**

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

390 **Cross-validation framework for single-trial GLM**

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

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

416 **GLMsingle algorithm**

417 Having described the essential aspects of the estimation framework above, we now turn to the steps in
418 the GLMsingle algorithm. GLMsingle involves fitting several different GLM variants. Each variant

419 includes polynomial regressors to characterize the baseline signal level: for each run, we include
420 polynomials of degrees 0 through $round(L/2)$ where L is the duration in minutes of the run.

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

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

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

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

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

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

455 To give a sense of the computational requirements of GLMsingle, we report here results for an example
456 scenario. We ran the MATLAB version of GLMsingle with default parameters on the first NSD scan
457 session for subj01 (1.8-mm standard-resolution version of the data). The scan session involved 750
458 trials and a data dimensionality of (81 voxels \times 104 voxels \times 83 voxels) = 699,192 voxels and (12
459 runs \times 226 volumes) = 2,712 time points. The code was run on an 32-core Intel Xeon E5-2670 2.60
460 GHz Linux workstation with 128 GB of RAM and MATLAB 9.7 (R2019b). The data were loaded in

461 single-precision format, resulting in a base memory usage of 8.4 GB of RAM (the data alone occupied
462 7.6 GB). Code execution (including figure generation and saving results to disk) took 4.8 hours (average
463 of 2 trials). The maximum and mean memory usage over the course of code execution was 38.0 GB
464 and 18.5 GB of RAM, respectively.

465 **GLMsingle outputs**

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

473 **Flexibility of GLMsingle**

474 Although GLMsingle provides default settings for the parameters that control its operation, the toolbox
475 is flexible and allows the user to adjust the parameters if desired. Modifying the parameters allows the
476 user to achieve a range of different behaviors, such as expanding the HRF library to include additional
477 candidate HRFs; changing the maximum number of nuisance regressors tested during GLMdenoise
478 (default is 10); modifying the range of shrinkage fractions evaluated for ridge regression (default is
479 0.05 to 1 in increments of 0.05); and running different flavors of GLM models that omit HRF fitting,
480 GLMdenoise, and/or ridge regression. For complete documentation, please refer to the GLMsingle
481 function descriptions and example scripts available at glmsingle.org.

482 **Application of GLMsingle to NSD and BOLD5000**

483
484 In order to assess the efficacy of GLMsingle for large-scale fMRI datasets, we tested GLMsingle on
485 the NSD (Allen et al., 2022) and BOLD5000 (Chang et al., 2019) datasets. Both datasets involve
486 presentation of many thousands of natural images. NSD and BOLD5000 share an overlapping subset of
487 stimuli from the Microsoft Common Objects in Context (COCO) database (Lin et al., 2014), enabling
488 direct comparison between the brain responses observed in the two datasets. However, there are a
489 number of differences between the datasets: the two datasets were collected at different field strengths,
490 with different event timings, and at different spatial and temporal resolution. In addition, while NSD
491 contains many repeated stimuli within each scan session, BOLD5000 contains very few. As such,
492 processing BOLD5000 requires grouping of input data across scan sessions to facilitate the cross-
493 validation procedures used in GLMsingle. This challenging processing scheme with respect to image
494 repetitions provides a strong test of the robustness of the GLMsingle technique.

495 **NSD Dataset**

496 For complete details of the NSD study, including scanning parameters, stimulus presentation, and
497 experimental setup, refer to the *Methods* section of the corresponding dataset paper (Allen et al., 2022).
498 In brief, a total of 8 subjects participated in the NSD experiment, each completing between 30-40
499 functional scanning sessions. For the full experiment, 10,000 distinct images from the Microsoft COCO
500 dataset were designed to be presented 3 times each over the course of 40 sessions. For computational
501 convenience and to make comparisons across subjects easier, only the first 10 NSD sessions from
502 subjects 1–4 are used for the analyses contained in this manuscript. Functional data were collected at
503 7T, with 1.8-mm isotropic resolution, and with a TR of 1.6 s. Each trial lasted 4 s, and consisted of the
504 presentation of an image for 3 s, followed by a 1-s gap. A total of 12 NSD runs were collected in one
505 session, containing either 62 or 63 stimulus trials each, for a total of 750 trials per session.

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

512 **BOLD5000 Dataset**

513 For complete details of the BOLD5000 study and methodology, refer to the corresponding dataset paper
514 ([Chang et al., 2019](#)). A total of 4 subjects participated in the BOLD5000 dataset (CSI1-4). A full dataset
515 contained 15 functional scanning sessions; subject CSI4 completed only 9 sessions before withdrawing
516 from the study. BOLD5000 involved presentation of scene images from the Scene UNderstanding
517 (SUN) ([Xiao et al., 2010](#)), COCO ([Lin et al., 2014](#)), and ImageNet ([Deng et al., 2009](#)) datasets. A total
518 of 5,254 images, of which 4,916 images were unique, were used as the experimental stimuli. 112 of the
519 4,916 distinct images were shown four times and one image was shown three times to each subject.
520 Functional data were collected at 3T, with 2-mm isotropic resolution, and with a TR of 2 s. Each trial
521 lasted 10 s, and consisted of the presentation of an image for 1 s, followed by a 9-s gap. A total of either
522 either 9 or 10 runs were collected in one session, containing 37 stimulus trials each, for a total of either
523 333 or 370 trials per session.

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

528 Because GLMsingle requires condition repetitions in order to perform internal cross-validation proce-
529 dures, and because BOLD5000 contains a limited number of within-session repetitions, we concatenated
530 data from groups of 5 sessions together before processing using GLMsingle. To account for differences
531 in BOLD signal intensity across different sessions, we performed a global rescaling operation to the
532 data within each session to roughly equate the time-series mean and variance across the 5 sessions
533 comprising one batch of data. Specifically, we first computed the global mean fMRI volume across all
534 5 sessions, and then, for each session, computed a linear fit between the mean volume from a single
535 session and the global mean volume. This yielded a multiplicative scaling factor applied to each session
536 in order to roughly equate signal intensities across sessions.

537 **Applying GLMsingle to NSD and BOLD5000**

538 We used GLMsingle to estimate single-trial BOLD responses in the NSD and BOLD5000 datasets.
539 For NSD, GLMsingle was applied independently to each scan session. For BOLD5000, groups of
540 5 sessions were processed together, following the rescaling procedure described above. The default
541 GLMsingle parameters were used for processing both NSD and BOLD5000, except that we evaluated
542 up to 12 nuisance regressors in GLMdenoise (default: 10).

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

551 Step 5, and reflects the additional use of ridge regression. For comparisons between GLMsingle and
552 Least-Squares Separate (LSS) signal estimation (**Figure 3**), 4 auxiliary beta versions were computed.
553 LSS betas were compared to those estimated using fractional ridge regression in the scenario of using
554 the canonical HRF (**AssumeHRF + LSS** vs. **AssumeHRF + RR**) and in the scenario of performing
555 HRF optimization using the GLMsingle library (**FitHRF + LSS** vs. **FitHRF + RR**). Our validation
556 analyses involve comparing optimized GLMsingle betas (b_2, b_3, b_4) against those estimated using the
557 baseline GLM approach (b_1), and performing an 8-way comparison incorporating both b_1-b_4 and the
558 4 auxiliary beta versions used for comparisons with LSS. Prior to all analyses, the responses of each
559 voxel were z-scored within each experimental session in order to eliminate potential nonstationarities
560 arising over time, and to equalize units across voxels.

561 Assessing the impact of GLMsingle

562

563 Analysis of voxel reliability

564 *Computing test-retest reliability* – To compute reliability, we repeated the following procedure for
565 each beta version. We first extracted the betas from trials that correspond to repetitions of the same
566 stimuli (NSD: 3 instances per stimulus; BOLD5000: 4 instances for subjects CSI1-3, and 3 for CSI4).
567 For each voxel, this yielded a matrix of dimensions (repetitions x images). To compute reliability,
568 Pearson correlation was computed between the average voxel response profiles for each possible unique
569 split-half of the data. Therefore, in the case of 4 available repetitions, the reliability for a voxel was
570 the average of 3 correlation values, with image repetitions grouped as follows: $corr(mean(1, 2)$ vs.
571 $mean(3, 4)$); $corr(mean(1, 3)$ vs. $mean(2, 4)$); $corr(mean(1, 4)$ vs. $mean(2, 3)$). In the case of 3
572 repetitions, the reliability was the average of: $corr(mean(1, 2)$ vs. (3)); $corr(mean(1, 3)$ vs. (2));
573 $corr(mean(2, 3)$ vs. (1)).

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

580 *Composite voxel reliability scores* – In comparing different beta versions output by GLMsingle, we
581 sought to understand whether the optimizations tended to affect all voxels equally, or whether the impact
582 was mediated by voxel reliability. We therefore measured how different beta versions differed in our
583 key outcome metrics (e.g. mean voxel reliability) as a function of the reliability of included voxels. To
584 achieve fair comparisons, we ensured that the same groups of voxels were compared at each reliability
585 threshold across beta versions. We achieved this by computing composite voxel reliability scores: the
586 mean reliability value in each voxel over beta versions b_1-b_4 . We then subselected groups of voxels
587 by applying varying threshold levels to the composite reliability scores. For analyses incorporating
588 the 4 auxiliary beta versions, composite reliability scores were computed as the mean across all 8 beta
589 versions.

590 *Effect of reliability on beta quality* – To quantify the performance of different beta versions as a function
591 of voxel reliability, composite scores were thresholded at increasing values (from Pearson $r = -0.2$ to
592 0.6, in steps of 0.05) to determine the included voxels at each reliability level. At each threshold, we
593 computed the difference between the reliability achieved by a given beta version and the composite
594 reliability (i.e. the average across beta versions). This difference was averaged across voxels, producing

595 traces that reflect the relative quality of data from each beta version compared to the group average,
596 across different levels of voxel reliability (**Figure 2b**).

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

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

615 **Analysis of temporal autocorrelation**

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

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

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

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

650 *Effect of reliability on temporal autocorrelation* – The effect of temporal autocorrelation in GLM betas
651 may vary depending on the relative responsiveness of different voxels to the experimental stimuli.
652 As such, we repeated the autocorrelation analyses several times, varying the expanse of voxels that
653 were included. We again relied on the aggregate reliability scores (computed previously) as a measure
654 of voxel quality, which are the average voxel reliabilities taken across all the beta versions under
655 consideration. This avoids biasing the voxel selection procedure. In **Figure 4**, we compare temporal
656 autocorrelation trends arising from analysis of voxels at two different reliability thresholds ($r = 0$ and
657 $r = 0.3$).

658 **Analysis of between-subject representational similarity**

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

667 *Between-subject RSA correlations* – For comparisons between subjects across NSD and BOLD5000,
668 we identified a subset of 241 images that overlapped between BOLD5000 and the portion of NSD being
669 analyzed for this manuscript. Once overlapping images were identified, the corresponding GLM betas
670 for each version were isolated, and averaged over all available repetitions within subject (3 for NSD, 4
671 for BOLD5000). Then, we used Pearson dissimilarity ($1 - r$) to compute RDMs over the averaged
672 betas for each subject, in each dataset. To assess the impact of voxel reliability on cross-subject
673 RDM correlations, this procedure was repeated across a range of voxel reliability inclusion levels
674 $r = [-1, 0, 0.05, 0.1, 0.15, 0.2, 0.25]$, using the beta version-averaged aggregate reliability scores
675 computed previously. Voxels inside the nsdgeneral ROI were used in this analysis. Once RDMs
676 were computed for each subject, using responses from the sets of stimuli detailed above, within- and
677 across-dataset RSA correlations were computed using the vectorized lower-triangular portions of each
678 RDM (**Figure 5b**).

679 **Analysis of MVPA decoding accuracy**

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

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

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

705 **Acknowledgments**

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

709 **Author Contributions**

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

714 **Conflict of Interest Statement**

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

717 References

718 Abdulrahman, H. and Henson, R. N. (2016). Effect of trial-to-trial variability on optimal event-related
719 fmri design: Implications for beta-series correlation and multi-voxel pattern analysis. *NeuroImage*,
720 125:756–766.

721 Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B.,
722 Pestilli, F., Charest, I., et al. (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and
723 artificial intelligence. *Nature neuroscience*, 25(1):116–126.

724 Bai, B. and Kantor, P. (2007). A shape-based finite impulse response model for functional brain images.
725 In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages
726 440–443. IEEE.

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

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

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

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

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

737 Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Hum Brain Mapp*, 8(2–
738 3):109–114.

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

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

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

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

749 D’Isanto, A. and Polsterer, K. L. (2018). Photometric redshift estimation via deep learning-generalized
750 and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy & Astrophysics*,
751 609:A111.

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

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

757 Handwerker, D. A., Gonzalez-Castillo, J., D'esposito, M., and Bandettini, P. A. (2012). The continuing
758 challenge of understanding and modeling hemodynamic variation in fmri. *Neuroimage*, 62(2):1017–
759 1023.

760 Handwerker, D. A., Ollinger, J. M., and D'Esposito, M. (2004). Variation of bold hemodynamic
761 responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*,
762 21(4):1639–1651.

763 Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Dis-
764 tributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*,
765 293(5539):2425–2430.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

796 Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L., and DiCarlo, J. J.
797 (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior
798 temporal cortex face processing network. *bioRxiv*.

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

801 Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L.
802 (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*,
803 pages 740–755. Springer.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

844 Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*,
845 5:399–426.

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

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

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

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

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

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