UC Berkeley

Title

Improved inference in coupling, encoding, and decoding models and its consequence for neuroscientific interpretation

Permalink

Authors

Sachdeva, Pratik S
Livezey, Jesse A
Dougherty, Maximilian E
et al.

Publication Date

DOI

Copyright Information

Improved inference in coupling, encoding, and decoding models and its consequence for neuroscientific interpretation

Pratik S. Sachdeva [a,b,c], Jesse A. Livezey [a,c], Maximilian E. Dougherty [c], Bon-Mi Gu [d], Joshua D. Berke [d,e], Kristofer E. Bouchard [a,c,f,g]

[a] *Redwood Center for Theoretical Neuroscience, University of California, Berkeley, 94720, CA, USA*
[b] *Department of Physics, University of California, Berkeley, 94720, CA, USA*
[c] *Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, 94720, CA, USA*
[d] *Department of Neurology, University of California, San Francisco, San Francisco, 94143, CA, USA*
[e] *Department of Psychiatry; Neuroscience Graduate Program; Kavli Institute for Fundamental Neuroscience; Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, 94143, CA, USA*
[f] *Computational Resources Division, Lawrence Berkeley National Laboratory, Berkeley, 94720, CA, USA*
[g] *Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, 94720, CA, USA*

ARTICLE INFO

ABSTRACT

Background: A central goal of systems neuroscience is to understand the relationships amongst constituent units in neural populations, and their modulation by external factors, using high-dimensional and stochastic neural recordings. Parametric statistical models (e.g., coupling, encoding, and decoding models), play an instrumental role in accomplishing this goal. However, extracting conclusions from a parametric model requires that it is fit using an inference algorithm capable of selecting the correct parameters and properly estimating their values. Traditional approaches to parameter inference have been shown to suffer from failures in both selection and estimation. The recent development of algorithms that ameliorate these deficiencies raises the question of whether past work relying on such inference procedures have produced inaccurate systems neuroscience models, thereby impairing their interpretation.
New method: We used algorithms based on Union of Intersections, a statistical inference framework based on stability principles, capable of improved selection and estimation.
Comparison: We fit functional coupling, encoding, and decoding models across a battery of neural datasets using both UoI and baseline inference procedures (e.g., ℓ_1-penalized GLMs), and compared the structure of their fitted parameters.
Results: Across recording modality, brain region, and task, we found that UoI inferred models with increased sparsity, improved stability, and qualitatively different parameter distributions, while maintaining predictive performance. We obtained highly sparse functional coupling networks with substantially different community structure, more parsimonious encoding models, and decoding models that relied on fewer single-units.
Conclusions: Together, these results demonstrate that improved parameter inference, achieved via UoI, reshapes interpretation in diverse neuroscience contexts.

1. Introduction

Neuroscience is undergoing a rapid growth in the size and complexity of experimental and observational data (Sejnowski et al., 2014; Marx, 2013). Realizing the benefits of these advances in data acquisition requires improvements in the statistical models characterizing the data, as well as the inference procedures used to fit those models (Stevenson and Kording, 2011; Bzdok and Yeo, 2017). For example, generalized linear models are appealing because the model parameters can be interpreted to gain insight into the underlying biological processes that generated the data (Stevenson et al., 2008; Okatan

et al., 2005; Truccolo et al., 2005; Pillow et al., 2008). However, even for this ubiquitously used class of models, the impact of an inference procedure's statistical properties on neurobiological interpretation is poorly appreciated.

These issues are particularly salient in systems neuroscience, where parametric models are often used to understand how neural activity is modulated by external factors (e.g., stimuli or a behavioral task) and internal factors (e.g., other neurons) (Paninski et al., 2007; Kass et al., 2018). The fitted parameter values, therefore, specify which factors are important in modulating neural activity, and how important they are. The specific relationships that a parametric model describes ultimately
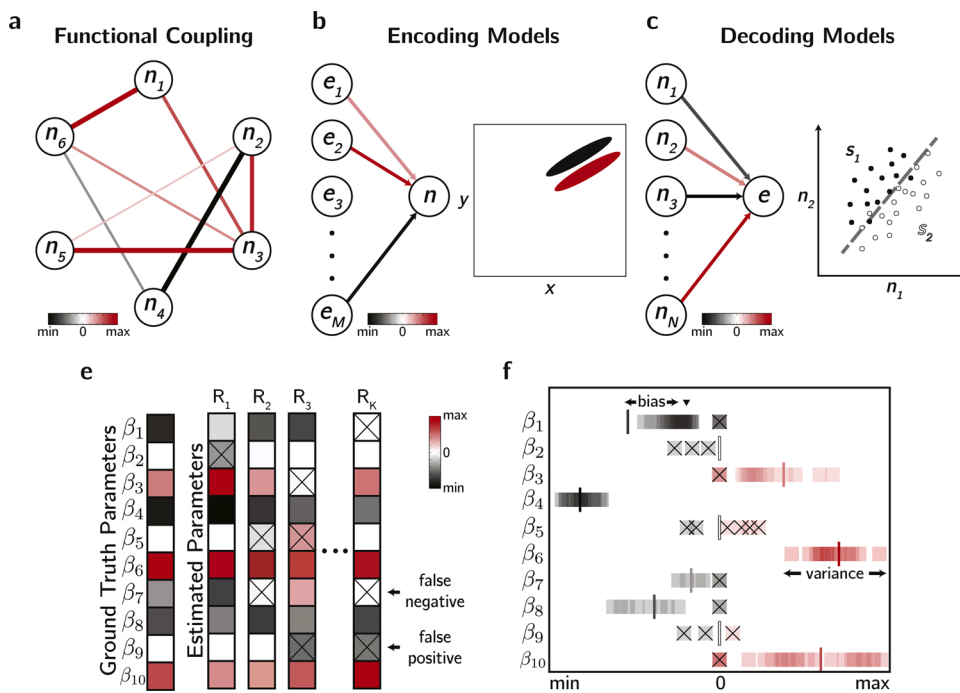
Fig. 1. Parametric models and statistical inference in systems neuroscience. (a–c) Three examples of parametric models widely used in systems neuroscience. (a) Functional coupling models characterize the statistical relationships between neurons in a population. (b) Encoding models map M internal or external factors $\{e_i\}_{i=1}^{M}$ to a neuronal response n. (c) Decoding models map the activities of N neurons in a population $\{n_i\}_{i=1}^{N}$ to an internal or external factor e. (e-f) Quantification of statistical selection and estimation performance. (e) The ground truth values of the parameters in an example model, given by the first column, along with estimated values across K different resamples of the data, denoted by $R_1, R_2, ..., R_K$. (f) The distribution of estimated values for each parameter in the ground truth model of the previous panel, with the true values denoted by vertical lines. For β_1, ▼ denotes the mean estimated value across resamples. False positives and false negatives are denoted with an ×.

frames how the model will be interpreted in a neuroscientific context, emphasizing the importance of accurate parameter inference.

For example, functional coupling models (Fig. 1a) capture the statistical dependencies between different functional units in the brain, at scales ranging from single units to functional areas (Stevenson et al., 2008, 2012; Okatan et al., 2005; Truccolo et al., 2005; Pillow et al., 2008; Babadi et al., 2010; Zhao et al., 2012; Song et al., 2013). These models can be used to construct networks (Bassett and Sporns, 2017; Bullmore and Sporns, 2009), which are analyzed with an assortment of tools from graph theory to characterize the population (Bassett and Bullmore, 2006; Barabási and Albert, 1999). Additionally, functional coupling networks are related to structural connectivity (Melozzi et al., 2019), used to assess directed influence amongst neurons (i.e., effective connectivity) (Friston et al., 2013; Seth et al., 2015), or related to external factors such as behavior, genetics, aging, or psychiatric conditions (Kiani et al., 2014; Fulcher and Fornito, 2016; Arnemann et al., 2018). Encoding models map the dependence of a brain signal (e.g., neuronal spikes) on external factors, such as stimuli (Fig. 1b) (Dayan and Abbott, 2001; Schwartz et al., 2006; Triplett and Goodhill, 2019). An example encoding model is a spatio-temporal receptive field of a visual cortex neuron, which maps the image space to the neuronal response (Fig. 1b, right) (Sharpee, 2013; Theunissen et al., 2001; Hubel and Wiesel, 1962). More complex encoding models of neural population data can be used to test theoretical and computational theories of neural coding (Rao and Ballard, 1999; Vinje and Gallant, 2000; Zhu and Rozell, 2013). On the other hand, decoding models map brain signals to external factors, using the activities in, e.g., a neural population, to predict a stimulus or task-relevant behavioral condition (Fig. 1c) (Glaser et al., 2017; Naselaris et al., 2011; Holdgraf et al., 2017; Bouchard et al., 2013; Bouchard and Chang, 2014; Livezey et al., 2019; Anumanchipalli et al., 2019). A common linear decoding model is the extraction of a hyperplane in the neural activity space, which provides a decision boundary for one of two behavioral conditions or stimuli (Fig. 1c, right: s_1, s_2) (Kiani et al., 2014; Yamashita et al., 2008; Rich and Wallis, 2016). Recent work has explored more complex decoders, using artificial neural networks (Livezey et al., 2019) or predictive latent representations (Pandarinath et al., 2018). Using decoding models for brain–computer interfaces has both clinical uses and scientific implications for understanding learning and motor control (Wander et al., 2013; Carmena

et al., 2003). Since these models are used to make scientific conclusions about the function of the brain, understanding the stability, accuracy, and parsimony of the inference procedures and resulting models is of paramount importance.

The utility of parametric models hinges on the assumption that the inference procedure used to fit them selects the correct parameters (i.e., specified as zero or non-zero) and properly estimates their values. The statistical consequences of improper selection are false positives or false negatives (Fig. 1d), while poor estimation results in high bias (Fig. 1e: e. g., β_1) or high variance (Fig. 1e: e.g., β_6). The neuroscientific consequences of statistical inference lie in the interpretation of the fitted parametric model. Selection informs which internal and external factors are relevant for predicting neural activity, and estimation specifies their relative importance. Importantly, accurate selection is not a natural byproduct of predictive capacity, as cross-validated predictive accuracy is often a poor criterion for feature selection. Specifically, model selection by held-out cross-validation predictive accuracy lacks guarantees on consistency and has been implicated in producing false positives (Shao, 1993, 1997; Zhang et al., 2010; Wasserman and Roeder, 2009; Liu et al., 2010). Thus, validating that an inference procedure can reliably select and estimate a model's parameters is vital to ensure that they motivate correct conclusions about neural activity.

These issues imply that, when fitting parametric models in a scientific context, multiple goals beyond predictive performance must be balanced to produce a scientifically meaningful model. In particular, achieving a parsimonious model, which uses the fewest number of features to sufficiently predict the response variable (i.e., finding the "simplest"), has long served as a goal in statistical model selection (Schwarz, 1978). One approach to model parsimony relies on the imposition of sparsity during feature selection, which has the added benefit of identifying a small subset of predictive features, facilitating the interpretability of the model (Tibshirani, 1996; Hastie et al., 2015). This is particularly relevant in high-dimensional settings where there are few task-relevant features and strong priors from domain knowledge for selection may not exist. Another desired property is stability, or the reliability of an inference algorithm when its inputs are slightly perturbed (Yu et al., 2013; Bousquet et al., 2002). For a model to be interpretable, its parameters must be robust to the often noisy processes that generated the data. Thus, encouraging stability in a model's
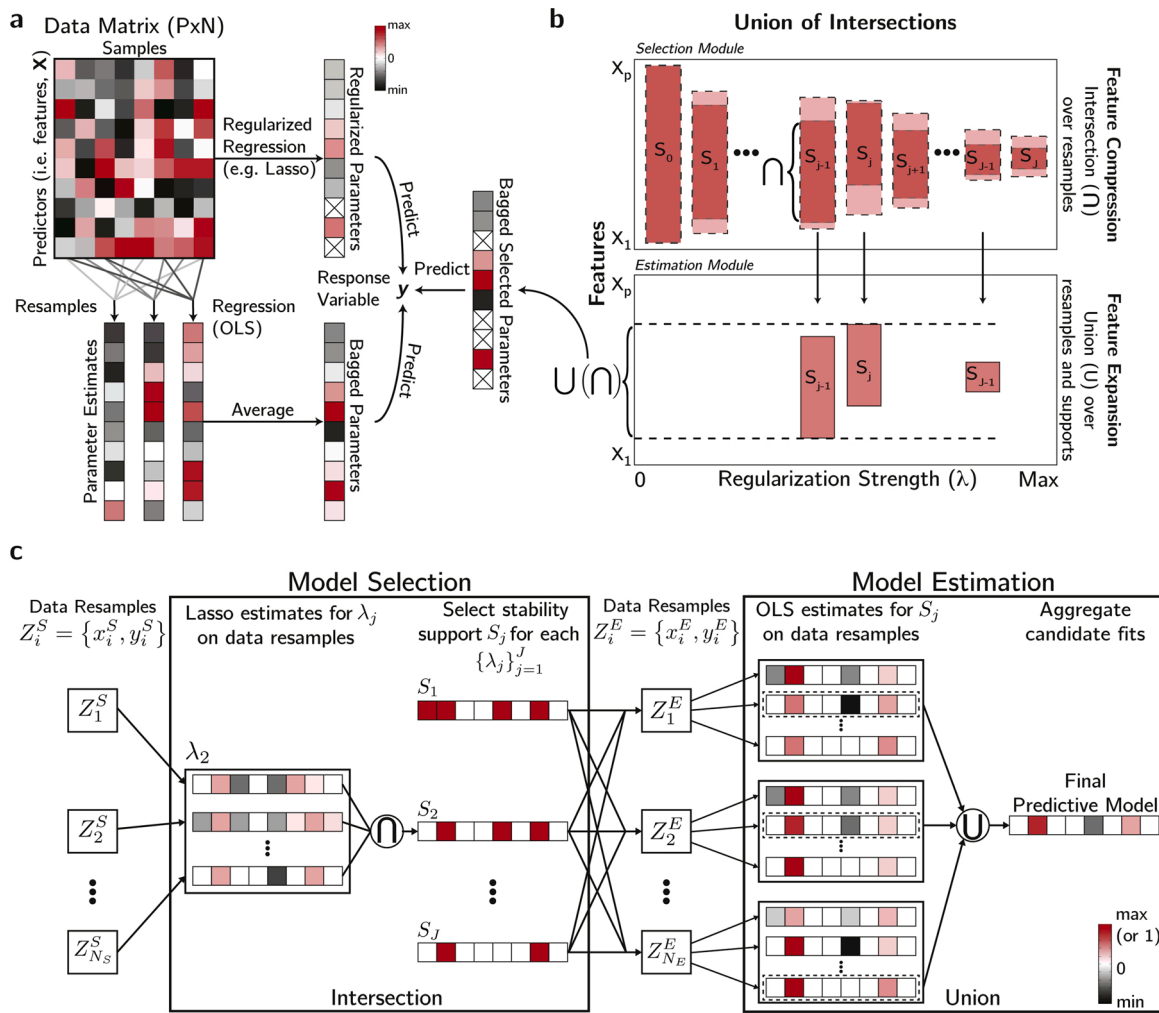
Fig. 2. The Union of Intersections framework combines ensemble and regularization approaches in model inference. (a) Schematic of regularization and ensemble methods. Top: Regularization can be used to perform feature selection. Features set exactly to zero are denoted by ×. Bottom: Ensemble procedures aggregate model fits across resamples of the data. (b) Schematic of the UoI framework. The *x*-axis corresponds to the set of regularization parameters while the *y*-axis corresponds to the feature space $(X_1, X_2, ..., X_p)$. Pink bands denote features included in a support or estimated model. Dark pink bands denote features included after the intersection step. (c) UoI$_{\text{Lasso}}$ depicted in a data-distributed fashion. *Model Selection (left):* Left column depicts the N_S data resamples, Z_i^S, during selection. Lasso fits are obtained for each resample, at different λ_j (e.g., left column: λ_2). Right column depicts the intersected supports S_j for each λ_j, with S_2 referring specifically to the estimates in the left column. Red here denotes 1, rather than maximum value. *Model Estimation (right):* Middle column denotes each data resample, Z_i^E, for estimation. Each resample is fit using each support $\{S_j\}_{j=1}^{J}$ (arrows before left column), generating a set of fits per resample (left column). The fit that achieves the best predictive performance for a given resample is chosen as that resample's candidate fit (left column: dashed lines). The candidate fits are aggregated to produce the final predictive model (right column).

parameter inference procedure will ensure that the features describing the relevant signal are selected and their correct contributions are properly estimated (Lim and Yu, 2016; Baldassarre et al., 2017). Until recently, inference procedures that sufficiently balanced selection and estimation, predictive performance, and stability were lacking. This raises the question of whether the usage of traditional inference procedures in systems neuroscience has adversely impacted neuroscientific interpretation and data-driven discovery.

Our recently introduced Union of Intersections (UoI) is an inference framework based on stability principles which enhances inference in a variety of common parametric models (Bouchard et al., 2017). The properties characterizing UoI models — sparsity, stability, and predictive accuracy — are well-suited to data-driven discovery in neuroscience, due to the high dimensionality and many sources of variability in these datasets. Furthermore, UoI is a frequentist approach, similar to the predominant traditional approaches used by neuroscientists (though we note recent development of Bayesian inference algorithms that also perform well in these settings (Chen et al., 2010; Zhou et al., 2016;

Huang and Tran, 2018)). Thus, we used UoI to assess whether common approaches to parameter inference in models are susceptible to improper feature selection and estimation, and if so, assess the consequences for model interpretability in a neuroscience context.

In this work, we used the UoI framework to fit functional coupling, encoding, and decoding models to diverse neural data in an effort to elucidate the impacts of precise selection and estimation on neuroscientific interpretation. We found that, compared to baseline procedures (e.g., ℓ_1-regularization), we obtained models with enhanced sparsity, improved stability, and significantly different parameter distributions, while maintaining predictive performance across recording modality, brain region, and task. Specifically, we obtained highly sparse coupling models of rat auditory cortex, macaque V1, and macaque M1 without loss in predictive performance. These models were used to construct functional networks that exhibited enhanced modularity and decreased small-worldness. We built parsimonious encoding models of mouse retinal ganglion cells and rat auditory cortex that more tightly matched with theory. These models were able to predict held-out neural re-

sponses with parameters that were as simple as possible, but no simpler. Lastly, we decoded task-relevant external factors from rat basal ganglia activity using fewer single units than baseline models. Overall, by utilizing improved inference algorithms during the fitting of parametric neural models, we constructed more sparse and stable models. We assessed the neuroscientific consequences of using these models, finding notable changes in secondary analyses.

2. Methods

Our goal is to demonstrate how the statistical properties of inference algorithms impact the fitting and interpretation of diverse parametric models commonly used in neuroscience. The main tools we use for this purpose are algorithms based on the Union of Intersections framework (Bouchard et al., 2017; Ubaru et al., 2017; Sachdeva et al., 2019). Thus, we organize the Methods as follows. First, in Section 2.1, we introduce the Union of Intersections framework, while providing other relevant background. Second, in Section 2.2, we describe a large-scale synthetic experiment comparing a specific UoI algorithm, UoI$_{\text{Lasso}}$, versus other algorithms on a synthetic dataset to motivate UoI's usage on neural datasets. Third, in Section 2.3, we describe the neural datasets on which we performed the model-fitting and subsequent analyses. Lastly, in Section 2.4, we provide the details of those analyses, outlining how model-fitting was performed for each coupling, encoding, and decoding model. We provide further details on subsequent analyses performed on the fitted models, such as statistical tests and network construction for coupling models. The code used to produce all figures and analyses discussed in this paper is publicly available on Github[1].

2.1. The Union of Intersections framework balances sparsity, stability, and predictive performance

Union of Intersections (UoI) is not a single method or algorithm, but a flexible framework into which other algorithms can be inserted for enhanced inference. In this work, we apply the UoI framework to generalized linear models, focusing on linear regression (UoI$_{\text{Lasso}}$), Poisson regression (UoI$_{\text{Poisson}}$) and logistic regression (UoI$_{\text{Logistic}}$). We refer the reader to UoI variants of other procedures, such as non-negative matrix factorization (Ubaru et al., 2017) and column subset selection (Bouchard et al., 2017).

Consider the general problem of mapping a set of p features $\mathbf{x} \in \mathbb{R}^{p \times 1}$ to a response variable $y \in \mathbb{R}$, of which we have N samples $\{\mathbf{x}_i, y_i\}_{i=1}^N$. For convenience, we focus on linear models, which require estimating p parameters $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ that linearly map $\mathbf{x}_i$ to y_i. We describe the UoI framework in this context, which involves the algorithm UoI$_{\text{Lasso}}$. The steps we detail, however, extend naturally to other penalized generalized linear models (Friedman et al., 2010). Typically, the mapping in linear models is corrupted by i.i.d. Gaussian noise ε:

$$y = \boldsymbol{\beta}^T \mathbf{x} + \varepsilon. \tag{1}$$

The parameters $\boldsymbol{\beta}$ can be inferred by optimizing the traditional least squares error on y:

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^N (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2, \tag{2}$$

where i indexes the N data samples. The UoI framework combines two techniques — regularization and ensemble methods — to balance sparsity, stability, and predictive performance, thereby improving on the traditional least squares estimate (Fig. 2a).

Structured regularization, or the inclusion of penalty terms in the objective function to restrict the model complexity, can be useful when a

subset of the β_i are exactly equal to zero, i.e., $\boldsymbol{\beta}$ is sparse. Sparsity implies that some features are not relevant for predicting the response variable. This assumption is often useful for data-driven discovery in biological settings, particularly for framing the interpretation of the model in the context of physical processes that generated the data. The identification of which β_i are non-zero can be viewed as a feature selection (or more generally, model selection) problem (Hastie et al., 2015). A common regularization penalty used for feature selection is the lasso penalty $|\boldsymbol{\beta}|_1$, or the ℓ_1-norm applied to the parameters (Tibshirani, 1996). For the case of linear regression, this creates an optimization problem of the form

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^N (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 + \lambda |\boldsymbol{\beta}|_1. \tag{3}$$

Solving Eq. (3) returns parameter estimates with some sparsity, provided that λ is appropriately chosen (Fig. 2a, top). Typically, λ, the degree to which feature sparsity is enforced, is unknown and must be determined through cross-validation or a penalized score function such as the Bayesian information criterion (BIC) (Schwarz, 1978) across a set of J hyperparameters $\{\lambda_j\}_{j=1}^J$. Importantly, solving the lasso problem simultaneously performs model selection (identifying the non-zero features) and model estimation (determining the specific values of those parameters). However, the application of the lasso penalty suffers from shrinkage (Tibshirani, 1996), or a parameter bias that erroneously reduces the magnitudes of the parameters (Fig. 2a, top: compare opacity of parameter estimates), and often does not correctly identify the true non-zero parameters (Fig. 2a, top: false positives).

On the other hand, ensemble procedures (e.g., bagging and boosting (Breiman, 1996; Friedman, 2001)) aggregate model fits across resamples of the data to improve the stability of parameter estimates (Fig. 2a, bottom). The more stable parameter estimates result in improved predictive accuracy. This is particularly desirable in biological settings, where model aggregation ensures that the relevant signal in noisy data is reflected in the parameter estimates. However, ensemble procedures do not perform feature selection.

UoI separates model selection and model estimation into two stages, with each stage utilizing ensemble procedures to promote stability. Specifically, model selection is performed through intersection (compressive) operations and model estimation through union (expansive) operations, in that order. This separation of parameter selection and estimation provides selection profiles that are robust and parameter estimates that have low bias and variance. Fig. 2b and 2c provide a visual depiction of the UoI framework, and A provides pseudocode for the UoI algorithm in generalized linear models. For UoI$_{\text{Lasso}}$, the procedure is as follows:

Model Selection. Define the support S as the set of non-zero parameters in an estimate $\widehat{\boldsymbol{\beta}}$. First, generate a regularization path of $\{\lambda_j\}_{j=1}^J$ spanning $(\varepsilon\lambda_{\max}, \lambda_{\max})$ where $\lambda_{\max}$ is analytically determined to result in an empty support and $\varepsilon = 10^{-3}$ (Friedman et al., 2010). For each λ_j, generate parameter estimates by solving the lasso optimization problem (Eq (3)) on N_S resamples of the data, and calculate a support for each resample-λ_j pairing. The intersection step requires that only the features that appear in a sufficient number of resamples are included in the final *stability support* S_j for λ_j. We depict this in Fig. 2b, top, where the light pink bands denote features that are included in the model due to regularization while the dark pink bands denote features that are included in the stability support after the intersection across resamples. The bands are arranged in order of increasing regularization strength (Fig. 2b: *x*-axis) and thus sparser (i.e., smaller) support sets (Fig. 2b: *y*-axis). Note that the stability support may be calculated with a hard intersection (e. g., Fig. 2c: Model Selection) or a soft intersection. In the former case, a feature must appear in the support of every resample to be included in S_j. In the latter, the feature must only appear in a sufficient fraction of supports which is a hyperparameter.
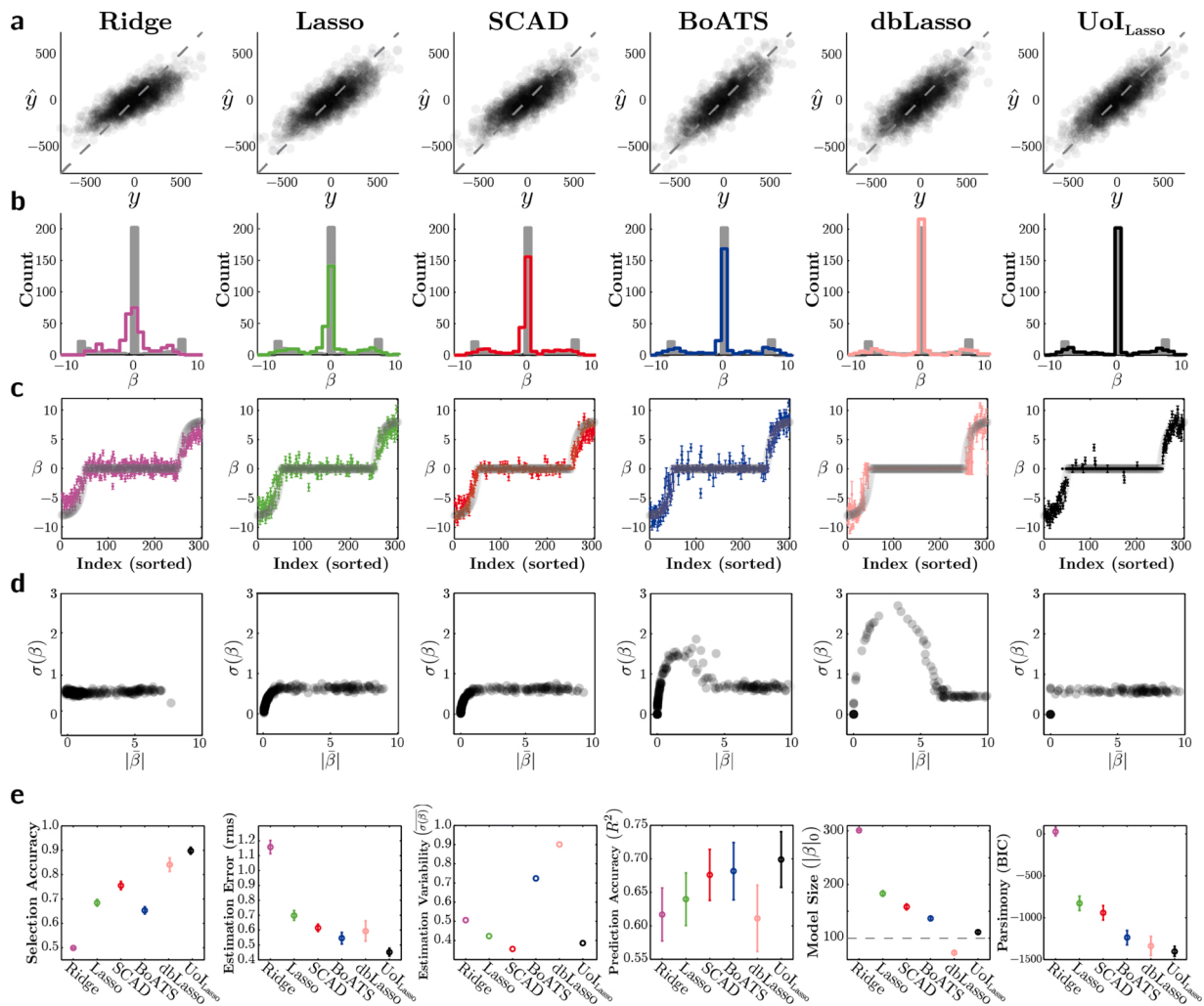

Fig. 3. UoI achieves superior selection and estimation performance on synthetic data over a battery of alternative inference algorithms. The performance of ridge regression (purple), lasso regression (green), smoothly clipped absolute deviation (SCAD; red), bootstrapped adaptive threshold (BoATS; blue), debiased lasso (dbLasso; coral), and UoI_{Lasso} (black) on data generated from a synthetic linear model. Panels a-d highlight separate measures per row, with each column referring to a separate inference algorithm. Each column in panel (e). directly compares a summary measure across inference algorithms. (a) Comparison of the predicted and true values of the response variable for held-out data. (b) Histogram of the estimated parameters (colored outline) compared to the distribution of true parameters (gray). (c) Estimated parameters (colored points; error bars denote the IQR across data resamples) compared to true parameters (gray), both sorted on the *x*-axis according to the value of the true parameter. (d) Variance of the parameter estimates across data resamples as a function of the mean estimated parameter's magnitude. (e) Direct comparison of the selection accuracy, estimation error (root mean square of parameter estimates), estimation variability (mean parameter variance), prediction accuracy (coefficient of determination), the inferred model's size (number of identified non-zero parameters, with the black dashed line denoting ground truth), and parsimony (Bayesian information criterion). Error bars denote IQR across data resamples.

Model Estimation. Generate N_E resamples of the data, and perform an unregularized fit on each resample using each support set S_j. For each resample, the support generating the fit that performs the best according to some metric is chosen as the "candidate fit" for that resample (Fig. 2b: bottom). We used the Bayesian information criterion (BIC) as this metric, which balances both predictive accuracy and model size (see Section 2.4.1 for more details on this choice). Unique supports may have the best performance across multiple resamples (e.g., only three unique supports, S_{j-1}, S_j, and S_{m-1}, are included for model averaging in Fig. 2b). The N_E candidate fits across resamples are unionized according to some metric (e.g., median, mean, etc.), resulting in a final parameter estimate (Fig. 2b, bottom). We use the median during the union step because it is more stable than the mean from a selection perspective. To be clear, the median will result in a parameter estimate set exactly equal to zero if that parameter is equal to zero in at least a majority of candidate fits. In contrast, the mean will likely result in a non-zero parameter estimate if even a single parameter value is non-zero. Note that, in the context of (generalized) linear models, the bagging of model parameters performed

in the estimation procedure is equivalent to the bagging of model predictions. For UoI_{Lasso}, the estimation procedure consists of applying Ordinary Least Squares to each stability support and resample combination (Fig. 2c: Model Estimation).

UoI's modular approach to parameter inference capitalizes on the feature selection achieved by stability selection and the unbiased, low-variance properties of the bagged OLS estimator. Furthermore, UoI's novel use of model aggregating procedures within its resampling framework allows it to achieve highly sparse (i.e., only using features robust to perturbations in the data) and predictive (i.e., only using features that are informative) model fitting. Importantly, this is achieved without imposing an explicit prior on the model distribution, and without formulating a non-convex optimization problem. Since the optimization procedures across resamples can be performed in parallel, the UoI framework is naturally scalable, a fact that we have leveraged to facilitate parameter inference on larger datasets (Sachdeva et al., 2019). The application of UoI_{Lasso} in a data-distributed manner is depicted in Fig. 2c. In the selection module, the first column depicts lasso estimates

across data resamples for a particular choice of regularization parameter, all of which can be fit in parallel (Fig. 2c, Model Selection: left column). In the estimation module, OLS estimates are fit across resamples and supports, which can be done in parallel (Fig. 2c, Model Estimation: left column).

2.2. Evaluation of Union of Intersections on synthetic data

We evaluated UoI$_{\text{Lasso}}$'s abilities as an inference procedure by assessing its performance on synthetic data generated from a linear model. The performances of UoI and five other inference procedures are depicted in Fig. 3: UoI$_{\text{Lasso}}$ (black), ridge regression (purple) (Hastie et al., 2015), lasso (green) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD; red) (Fan and Li, 2001), bootstrapped adaptive threshold selection (BoATS; blue) (Bouchard, 2015), and debiased lasso (dbLasso; coral) (Javanmard and Montanari, 2014).

The linear model consisted of $p = 300$ total parameters, with $k = 100$ non-zero parameters (thereby having sparsity $1 - k/p = 2/3$). The non-zero ground truth parameters were drawn from a parameter distribution characterized by exponentially increasing density as a function of parameter magnitude (Fig. 3b: gray histograms). We used $N = 1200$ samples generated according to the linear model (1) with noise magnitude chosen such that $\text{Var}(\varepsilon) = 0.2 \times |\boldsymbol{\beta}|_1$. We report metrics according to their statistics across 100 randomized cross-validation samples of the data.

In Fig. 3a, we show scatter plots comparing the predicted and actual values of the observation variable on held-out data samples. We visualized how well the inference procedures captured the underlying parameter distribution by comparing the histograms of (average) estimated model parameters (colors) overlaid on the ground truth model parameters (grey) (Fig. 3b). We additionally plotted parameter bias and variance, first by comparing the mean estimated value ($\pm$ standard deviation) against the ground truth parameter value (Fig. 3c), and then examining the standard deviation of the parameter estimates as a function of their mean estimated value (Fig. 3d).

Fig. 3a-d captures the improvements that the UoI framework offers in parameter inference. UoI$_{\text{Lasso}}$ is designed to maximize prediction accuracy (Fig. 3a) by first selecting the correct features (Fig. 3b), and then estimating their values with high accuracy (Fig. 3c) and low variance (Fig. 3d). By separating model selection and model estimation, UoI$_{\text{Lasso}}$ benefits from strong selection (as in BoATS and debiased Lasso), but with the low variability of the structured regularizers (Lasso, SCAD), while alleviating shrinkage with its nearly unbiased estimates.

We quantified the performance of the inference algorithms on synthetic data using a variety of metrics capturing selection, bias, variance, and prediction accuracy. Specifically, these metrics were:

(4) **Selection Accuracy.** The selection accuracy, or set overlap, is a measure of how well the estimated support captures the ground truth support. Define S_β as the set of features in the ground truth support, $S_{\widehat{\beta}}$ as the set of features in the estimated model's support, $|S|$ as the cardinality of S, and Δ as the symmetric set difference operator. Then the selection accuracy is defined as

$$\text{selection accuracy}(S_\beta, S_{\widehat{\beta}}) = 1 - \frac{|S_\beta \Delta S_{\widehat{\beta}}|}{|S_\beta| + |S_{\widehat{\beta}}|}. \tag{4}$$

The selection accuracy is bounded in $[0,1]$, taking value 0 if $S_{\widehat{\beta}}$ and S_β have no elements in common, and taking value 1 iff they are identical.

(5) **Estimation error.** The estimation error of the p fitted parameters $\widehat{\beta}$, with ground truth parameters β, is defined as the root mean square error, or

$$\text{estimation error} = \sqrt{\frac{1}{p} \sum_{i=1}^{p} (\beta_i - \widehat{\beta}_i)^2}. \tag{5}$$

(6) **Estimation variability.** The estimation variability for parameter β_i is defined as the parameter standard deviation $\sigma(\beta_i)$. We calculated this quantity by taking the variance of the estimated parameter $\widehat{\beta}_i$ over R resamples of the data:

$$\sigma(\beta_i) = \sqrt{\frac{1}{R} \sum_{j=1}^{R} (\beta_i - \widehat{\beta}_{ij})^2}, \tag{6}$$

where j indexes the resample. To summarize this measure across all p parameters in a model, we took the average, i.e., $\sigma = \frac{1}{p}\sum_{i=1}^{p}\sigma(\beta_i)$.

(7) **Predictive performance.** To capture predictive performance, we used the coefficient of determination (R^2) evaluated on held-out data:

$$R^2 = 1 - \frac{\sum_{i=1}^{D}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{D}(y_i - \bar{y})^2} \tag{7}$$

where y_i is the ground truth response for sample i, $\widehat{y}_i$ its corresponding predicted value, and $\bar{y}$ the mean of the response variable over trials. R^2 has a maximum value of 1, when the model perfectly predicts the response variable across samples. R^2 values below zero indicate that the model is worse than an intercept model (i.e., simply using the mean value to predict across samples).

(8) **Model Parsimony.** We evaluated model parsimony using the Bayesian information criterion (BIC) (Schwarz, 1978):

$$\text{BIC} = k log(D) - 2 log \ell(\widehat{\beta}). \tag{8}$$

Here, D is the number of samples, k is the number of parameters estimated by the model, and $log\ell(\widehat{\beta}) = p(\widehat{\beta}|\mathcal{D}, m)$ is the log-likelihood of the parameters $\widehat{\beta}$ under data $\mathcal{D}$ and model m. Thus, the BIC includes a penalty that encourages models to be more sparse (first addend) while still accounting for predictive accuracy (second addend). Importantly, the BIC is evaluated on the data that the model was trained on (rather than held-out data). It is typically used as a model selection criterion (in lieu of, for example, cross-validation). When used as a model selection criterion, the model with lower BIC is preferred.

UoI$_{\text{Lasso}}$ generally resulted in the highest selection accuracy (Fig. 3e, first column), parameter estimates with lowest error (Fig. 3e, second column) and competitive variance (Fig. 3e, third column). In addition, it led to the best prediction accuracy (Fig. 3e, third column). UoI$_{\text{Lasso}}$ best captured the true model size (Fig. 3e, fourth column), avoiding the abundance of false positives suffered by most other inference algorithms. UoI$_{\text{Lasso}}$'s enhanced predictive performance with fewer features resulted in superior model parsimony (Fig. 3e, fifth column).

A robust set of experiments have been conducted comparing UoI$_{\text{Lasso}}$ to these methods in other settings. Specifically, these experiments assessed UoI$_{\text{Lasso}}$'s performance across a range of ground truth model sparsities, parameter distributions, and noise levels. Overall, UoI excels at parameter inference relative to other models across all these settings. These results can be found in the appendix of the original UoI paper (Bouchard et al., 2017).

Fig. 3 demonstrates the superiority of UoI methods over a battery of other approaches for a linear model. However, any generalized linear model fits within the UoI framework. For example, we used the Poisson (UoI$_{\text{Poisson}}$) and logistic (UoI$_{\text{Logistic}}$) variants of the UoI algorithm in this

study. Thus, we compared the performance of UoI to baseline procedures for Poisson and logistic regression. Specifically, we created similar synthetic datasets as described above, but in the Poisson and logistic contexts, and attempted to estimate the ground truth parameters using UoI and lasso-penalized Poisson and logistic regression. These results are detailed in B. We arrived at similar conclusions: UoI exhibits increased selection accuracy, decreased estimation error, comparable variability, improved prediction accuracy, and enhanced model parsimony relative to the baseline procedures. These results are in line with theoretical guarantees on support recovery in generalized linear models (Van de Geer et al., 2008; Bühlmann and Van De Geer, 2011).

2.3. Neural recordings

We sought to demonstrate impact of improved inference on parametric models across a diversity of datasets, spanning distinct brain regions, animal models, and recording modalities. We used micro-electrocorticography recordings obtained from rat auditory cortex (for coupling and encoding models), single-unit recordings from macaque visual and motor cortices (coupling models), single-unit recordings from isolated rat retina (encoding models), and single-unit recordings from basal ganglia (decoding models). We briefly describe the experimental and preprocessing steps for each dataset.

2.3.1. Recordings from auditory cortex
Auditory cortex (AC) data was comprised of cortical surface electrical potentials (CSEPs) recorded from rat auditory cortex with a custom fabricated micro-electrocorticography (μECoG) array. The μECoG array consisted of an 8×16 grid of 40 μm diameter electrodes. Anesthetized rats were presented with 50 ms tone pips of varying amplitude (8 different levels of attenuation, from 0 dB to -70 db) and frequency (30 frequencies equally spaced on a log-scale from 500 Hz to 32 kHz). Each frequency-amplitude combination was presented 20 times, for a total of 4200 samples. The response for each trial was calculated as the z-scored, to baseline, high-γ band analytic amplitude of the CSEP, calculated using a constant-Q wavelet transform. Of the 128 electrodes, we used 125, excluding 3 due to faulty channels. Data was recorded by Dougherty & Bouchard (DB). Further details on the surgical, experimental, and preprocessing steps can be found in Dougherty et al. (2019).

2.3.2. Recordings from primary visual cortex
We analyzed three primary visual cortex (V1) datasets, comprised of spike-sorted units simultaneously recorded in three anesthetized macaque monkeys. Recordings were obtained with a 10×10 grid of silicon microelectrodes spaced 400 μm apart and covering an area of 12.96 mm². Monkeys were presented with grayscale sinusoidal drifting gratings, each for 1.28 s. Twelve unique drifting angles (spanning $0°$ to $330°$) were each presented 200 times, for a total of 2400 trials per monkey. Spike counts were obtained in a 400 ms bin after stimulus onset. We obtained $[106, 88, 112]$ units from each monkey. The data was obtained from the Collaborative Research in Computational Neuroscience (CRCNS) data sharing website (Teeters et al., 2008) and was recorded by Kohn and Smith (KS) (Kohn and Smith, 2016). Further details on the surgical, experimental, and preprocessing steps can be found in Smith and Kohn (2008) and Kelly et al. (2010).

2.3.3. Recordings from primary motor cortex
Primary motor cortex (M1) data was comprised of spike-sorted units simultaneously recorded in the motor cortex of Rhesus macaque monkey. Recordings were obtained with a chronically implanted silicon microelectrode array consisting of 96 electrodes spaced at 400 μm and covering an area of 16 mm². We used three datasets, consisting of three recording sessions from monkey I. The behavioral task required the monkey to make self-paced reaches to targets arranged on a 8×17 grid.

Spike counts were binned at 150 ms over the course of the entire recording session, resulting in $[4089, 4767, 4400]$ samples per recording session. We obtained $[136, 146, 147]$ units from each dataset. Data was recorded by O'Doherty et al. (OCMS) and obtained from Zenodo (O'Doherty et al., 2017). Further details on the surgical, experimental, and preprocessing steps can be found in Makin et al. (2018).

2.3.4. Recordings from retina
Retina data comprised spiking activity, extracellularly recorded from isolated mice retina. Recordings were obtained using a 61-electrode array. Isolated retina were presented with a flicking black or white bar stimulus according to a pseudo-random binary sequence for a period of 16.6 ms. We utilized recordings from 23 different retinal ganglion cells. Data was obtained from CRCNS and recorded by Zhang et al. (2014). Further details on the surgical, experimental, and preprocessing steps can be found in Lefebvre et al. (2008).

2.3.5. Recordings from basal ganglia
Basal ganglia data comprised tetrode recordings from two regions of rat basal ganglia: the globus pallidus pars externa (GPe: 18 units) and substantia pars nigra reticulata (SNr: 36 units). Recordings were performed during a rodent stop-signal task. Briefly, a rat was prompted to enter a center port with a light cue. The rat remained in the port until a Go cue (audio stimulus at 1 kHz or 4 kHz) which directed a lateral head movement to the left or right ports. On a subset of trials, the Go cue was followed by a Stop signal (white noise burst), indicating that the rat should remain in the center port. We utilized the successful Go trials, in which the rat was not given a Stop signal and successfully entered the correct port (186 trials). We used the spike count in the first 100 ms after the rat exited the center port to predict the behavioral condition (left or right). Further details on the surgical, experimental, and preprocessing steps can be found in Gu et al. (2020).

2.4. Neural data analysis and model fitting

All models fit to neural data consisted of various generalized linear models, depending on the application. We trained all baseline models using either the `glmnet` (Friedman et al., 2010) or `scikit-learn` (Pedregosa et al., 2011; Buitinck et al., 2013) packages. Meanwhile, we trained all UoI models using the `pyuoi` package (Sachdeva et al., 2019). This section is organized as follows: first, we discuss model fitting, including details on the estimation module in UoI, followed by data analysis for coupling, encoding, and decoding models. Second, we detail model evaluation, including the measures used to assess each model, statistical tests, calculation of effect size, and the cross-validation approach for evaluation.

2.4.1. Model selection criterion in the estimation module
In the UoI framework, the estimation module operates by unionizing fitted stability supports across resamples. Thus, the module requires a criterion by which to choose the best fitted stability support per resample. A natural choice, akin to cross-validation, is the out-of-resample validation performance according to some measure (e.g., R^2, deviance, etc.). This is a principled approach when predictive accuracy is the only usage of the model. However, in this context, where we frame the fitted parameters of the model in the underlying neuroscience, parameter selection implicitly becomes an additional goal that must be reflected in the model criterion. Cross-validated predictive accuracy is often not sufficient in these cases. Variations of cross-validation have been shown to be model inconsistent (Shao, 1993, 1997), with theoretical guarantees on its probability of overfitting (Zhang et al., 2010). Furthermore, empirically, it has been shown to overfit, and suffer from false positives (Wasserman and Roeder, 2009; Liu et al., 2010). Therefore, we instead utilized the Bayesian information criterion in the estimation module for each model, which has been shown to be model selection consistent (Shao, 1993). Furthermore, the BIC is a principled

choice for model selection criterion because it can be couched as an approximation to Bayes factors (Neath and Cavanaugh, 2012a).

2.4.2. Data analysis for coupling models

We used UoI$_{\text{Lasso}}$ (rat auditory cortex) and UoI$_{\text{Poisson}}$ (macaque V1 and M1) to fit coupling models. The auditory cortex model can be described with a linear model as

$$n_i = \beta_{i0} + \sum_{j=1, j\neq i}^{p} \beta_{ij} n_j + \varepsilon \tag{9}$$

where n_i is the high-gamma activity of the *i*th electrode on a trial. The baseline procedure consisted of a lasso optimization with coordinate descent, while the UoI approach utilized UoI$_{\text{Lasso}}$. The model for the spiking datasets, which utilizes a Poisson generalized linear model, can be written as

$$\mu_i = exp\left(\beta_{i0} + \sum_{j=1, j\neq i}^{p} \beta_{ij} n_j\right), \tag{10}$$

$$n_i \sim \text{Poisson}(\mu_i). \tag{11}$$

where n_i corresponds to the spike count of the *i*th neuron. The corresponding objective function for this model is the log-likelihood,

$$\mathcal{L}_i\left(\boldsymbol{\beta} | \left\{n_1^k, \ldots, n_p^k\right\}_{k=1}^{D}\right) = \sum_{k=1}^{D}\left[n_i^k\left(\beta_{0i} + \sum_{j=1, j\neq i}^{p} n_j^k \beta_{ij}\right) - exp\left(\beta_{0i} + \sum_{j=1, j\neq i}^{p} n_j^k \beta_{ij}\right)\right] \tag{12}$$

where *i* denotes that this model corresponds to the *i*th neuron, *j* indexes over the remaining neurons, and *k* indexes over the *D* data samples. The baseline approach consisted of applying coordinate descent to solve this objective function with a lasso penalty:

$$\mathcal{L}_{i,\text{baseline}}\left(\boldsymbol{\beta} | \left\{n_1^k, \ldots, n_p^k\right\}_{k=1}^{D}\right) = \frac{1}{D}\mathcal{L}_i + \lambda_1 |\boldsymbol{\beta}|_1 \tag{13}$$

where λ_1 is a hyperparameter specifying the strength of the ℓ_1 penalty. Note that the intercept terms were not penalized. Meanwhile, in UoI$_{\text{Poisson}}$, we utilized the same objective function Eq (13) in the selection module. In the estimation module, we used Eq. (12) with a very small ℓ_2 penalty for numerical stability purposes. The specific optimization algorithm was a modified orthant-wise L-BFGS solver (Gong and Ye, 2015).

Data analysis for encoding models

For retinal data, we fit spatio-temporal receptive fields (STRFs) frame-by-frame. Specifically, the STRF was comprised of *F* frames $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_F$, each a vector of size *M* and spanning Δt seconds. For neuron *i* and frame *k*, the encoding model consisted of

$$n_i(t) = \beta_0 + \boldsymbol{\beta}_k^T \text{e}(t - k\Delta t) \tag{14}$$

where $n_i(t)$ is the spike count at timepoint *t* and $\text{e}(t - k\Delta t)$ is flicking bar stimulus value at *k* bins before *t*. We fit the *F* models using lasso (baseline) and UoI$_{\text{Lasso}}$, and created the final STRF by concatenating the parameter values $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_F]$

The tuning model for the rat auditory recordings was constructed using Gaussian basis functions. We used eight Gaussian basis functions spanning the log-frequency axis with means $\{\mu_j\}_{j=1}^{8}$ (Stevenson et al., 2008, 2012). Thus, the high-gamma activity n_i of electrode *i* in response to frequency *f* was

$$n_i(f) = \beta_{i0} + \sum_{j=1}^{8} \beta_{ij} exp\left(-\frac{logf - \mu_j}{2\sigma^2}\right) \tag{15}$$

We chose $\sigma^2 = 0.64$ octaves so that basis functions sufficiently spanned

the plane. We chose $p = 8$ basis functions because this was the minimum number of basis functions for which every electrode had a selection ratio less than 1. We fit Eq. (15) using cross-validated lasso as the baseline. To characterize the relationship between selection ratio and predictive performance of the rat AC tuning models, we fit trendlines across models using Gaussian process regression. Specifically, we utilized a regressor with radial basis function kernel (length scale $\ell = 0.01$) and a white noise kernel (noise level $\alpha = 0.1$).

Lastly, we fit encoding models for the macaque V1 and M1 datasets using cosine basis functions as a function of grating and reach angle, respectively (Stevenson et al., 2012). Importantly, we modeled the spike count of the *i*th neuron after a variance stabilizing square-root transform, which is typically used when a Gaussian model is applied to data exhibiting Poisson variability (Yu et al., 2008). For the M1 dataset, the encoding model can be written as

$$r_i = \beta_{i0} + \beta_{i1}cos(\theta) + \beta_{i2}sin(\theta) + \varepsilon \tag{16}$$

where $r_i = \sqrt{n_i}$ is the square-rooted spike count of the *i*th neuron, and θ is the angle of the reach (Stevenson et al., 2011). The encoding model for the V1 neurons was similar, aside from an adjustment in period:

$$r_i = \beta_{i0} + \beta_{i1}cos(2\theta) + \beta_{i2}sin(2\theta) + \varepsilon, \tag{17}$$

where θ is the angle of the grating. We adjusted the period of the basis functions because the single-units were responsive to gratings drifting in either direction along an axis. Since these models only use two basis functions, we fit them using ordinary least squares (OLS) with no regularization. We did not perform a comparison between the baseline model and a UoI-fitted model, since the OLS estimator is a consistent estimator. These encoding models were only used when examining the relationship between network and tuning structure.

Data analysis for decoding models

We fit decoding models to basal ganglia recordings as binary logistic regression models. The model expresses the probability of one experimental condition *e* (e.g., the rat entering the left port) as

$$\text{Pr}[e = \text{left}] = \text{sigmoid}\left(\beta_0 + \sum_{j=1}^{p} \beta_j n_j\right), \tag{18}$$

where n_i is the neural activity of the *i*th neuron. The corresponding objective function is the log-likelihood, or

$$\mathcal{L}_i\left(\boldsymbol{\beta} | \left\{n_1^k, \ldots, n_p^k\right\}_{k=1}^{D}\right) = \sum_{k=1}^{D}\left[log\left(1 + exp(\beta_0 + \sum_{j=1}^{p} \beta_j n_j^k)\right) - e^k(\beta_0 + \sum_{j=1}^{p} \beta_j n_j^k)\right] \tag{19}$$

The baseline approach consisted of solving this objective function with an ℓ_1 penalty. UoI$_{\text{Logistic}}$ utilized objective function (19) with an ℓ_1 penalty in the selection module, and Eq. (19) alone in the estimation module.

2.4.3. Network creation and analysis

We performed secondary analyses on the coupling models by constructing graphs using the fitted parameters. We then analyzed these networks with standard graph theoretic measures. Here, we detail how we constructed the networks and the measures we used to analyze them.

We created directed graphs by filling the adjacency matrix A_{ij} with the coefficient β_{ij} (i.e., the coupling coefficient for neuron *j* in the coupling model for neuron *i*). Meanwhile, we created undirected networks from coupling models by symmetrizing coefficients (Wang et al., 2014). Specifically, the symmetric adjacency matrix satisfies $A_{ij} = \frac{1}{2}\left(\beta_{ij} + \beta_{ji}\right)$ where β_{ij} is the coefficient specifying neuron *i*'s dependence on neuron *j*'s activity, and vice versa for β_{ji}. Thus, the network lacked an edge between vertices (neurons) *i* and *j* if only if neuron *i*'s coupling

model did not depend on neuron j, and neuron j's coupling model did not depend on neuron i. This adjacency matrix is weighted in that each entry depends on the magnitudes of the coupling coefficients. However, we can also consider an unweighted, undirected graph, whose adjacency matrix is simply the binarization of A_{ij}.

We analyzed the networks with the following measures:

- **In- and out-degree.** In a directed graph, the in-degree of a vertex is the number of incoming edges to that vertex. Meanwhile, the out-degree is the number of outgoing edges from the vertex. We examine the distribution of in-degrees and out-degrees across vertices in each group, which is dependent on the sparsity of the coupling models.

- **Modularity.** The modularity Q is a scalar value that measures the degree to which a network is divided into communities (Newman, 2006). We operate on the undirected graph described by the binarized adjacency matrix, described above. Suppose each vertex v is partitioned into one of c communities, where vertices within a community are more likely to be connected with each other than vertices between communities. Then, the modularity is defined as

$$Q = \frac{1}{2m} \sum_{v,w} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w) \qquad (20)$$

where c_v denotes community identity, k_v is the degree of vertex v, and m is the total number of edges. Thus, the modularity is greater than zero when there exist more edges between vertices within the same community than might be expected by chance according to the degree distribution. Specifically, Q is bounded within the range $[-1/2, 1]$, where $Q > 0$ indicates the existence of community structure. We calculated modularity with the Clauset–Newman–Moore greedy modularity maximization algorithm (Clauset et al., 2004). This procedure assigns vertices to communities by greedily maximizing the modularity, and then calculating Q using the ensuing community identities.

- **Small-worldness.** Small-world networks are characterized by a high degree of clustering with a small characteristic path length (Watts and Strogatz, 1998; Bassett and Bullmore, 2006, 2017). There are multiple measures used to quantify the degree to which a network is small-world. We use ω, which can be expressed as

$$\omega = \frac{L_r}{L} - \frac{C}{C_\ell}, \qquad (21)$$

where L is the characteristic path length of the network, L_r is the characteristic path length for an equivalent random network, C is the clustering coefficient, and C_ℓ is the clustering coefficient of an equivalent lattice network (Telesford et al., 2011). The quantity ω is bounded within $[-1, 1]$, where ω close to 0 indicates that the graph is small-world. When ω is close to 1, the graph is closer to a random graph, while ω close to -1 implies the graph is more similar to a lattice graph.

2.4.4. Model evaluation

We used the following measures to evaluate the models fit to neuroscience data:

(22) **Selection Ratio.** We evaluate the sparsity of estimated models with the selection ratio, or the fraction of parameters fitted to be non-zero:

$$\text{selection ratio} = \frac{k}{p}, \qquad (22)$$

where p is the total number of parameters available to the model

and k is the number parameters that a model-fitting procedure explicitly sets non-zero.

(23) **Predictive performance.** We utilized several measures of predictive performance, depending on the model. For linear models (i.e., a generalized linear model with an identity link function), we used the coefficient of determination (R^2) evaluated on held-out data, as detailed in Section 2.2 Recall that R^2 values below zero indicate that the model is worse than an intercept model (i.e., simply using the mean value to predict across samples).

For Poisson regression, or a generalized linear model with a logarithmic link function, we utilized the deviance, which is the difference in log-likelihood between the saturated model and the estimated model (Nelder and Wedderburn, 1972). The saturated model has parameters specifically chosen to reproduce the observed values. For the Poisson log-likelihood, the expression for the deviance as a function of the estimation parameters $\widehat{\beta}$ is given by

$$\text{deviance}(\widehat{\beta}) = \left[\sum_{i=1}^{D} y_i log(y_i) - y_i \right] - \left[\sum_{i=1}^{D} y_i (\widehat{\beta}_0 + \widehat{\beta}^T x_i) - exp(\widehat{\beta}_0 + \widehat{\beta}^T x_i) \right],$$

$$(23)$$

where $\{x_i, y_i\}_{i=1}^{D}$ denote the features and response variable of the model, respectively. Note that lower deviance is preferred, in contrast to the coefficient of determination. For logistic decoding models, we used the classification accuracy on held-out data as the measure of predictive performance.

(24) **Model Parsimony** As detailed in Section 2.2, we evaluated model parsimony using the Bayesian information criterion (BIC). Recall that the BIC includes a penalty that encourages models to be more sparse while still accounting for predictive accuracy. Importantly, the BIC is evaluated on the data that the model was trained on (rather than held-out data). It is typically used as a model selection criterion (in lieu of, for example, cross-validation). When used as a model selection criterion, the model with lower BIC is preferred.

(25) **Effect Size.** To fully capture the difference in model evaluation metrics beyond statistical significance, we measure effect size using Cohen's d (Cohen, 2013). For two groups of data with sample sizes D_1, D_2, means μ_1, μ_2, and standard deviations s_1, s_2, Cohen's d is given by

$$d = \left| \frac{\mu_1 - \mu_2}{s} \right| \qquad (24)$$

where s is the pooled standard deviation:

$$s = \sqrt{\frac{(D_1 - 1)s_1^2 + (D_2 - 1)s_2^2}{D_1 + D_2 - 2}}. \qquad (25)$$

We often considered cases where $D_1 = D_2$, implying that $s = \sqrt{\frac{s_1^2 + s_2^2}{2}}$. Values of d on the order of 0.01 indicate very small effect sizes, while $d > 1$ indicates a very large effect size (Sawilowsky, 2009).

(26) **Statistical Tests.** We used the Wilcoxon signed-rank test (Wilcoxon, 1992) to assess whether the distributions of selection ratios and predictive performances, across units, were significantly different between the UoI models and the baseline models. Importantly, we did not apply the test to the distribution of BICs, since differences in BIC are better interpreted as approximations to Bayes factors (Neath and Cavanaugh, 2012b). To assess
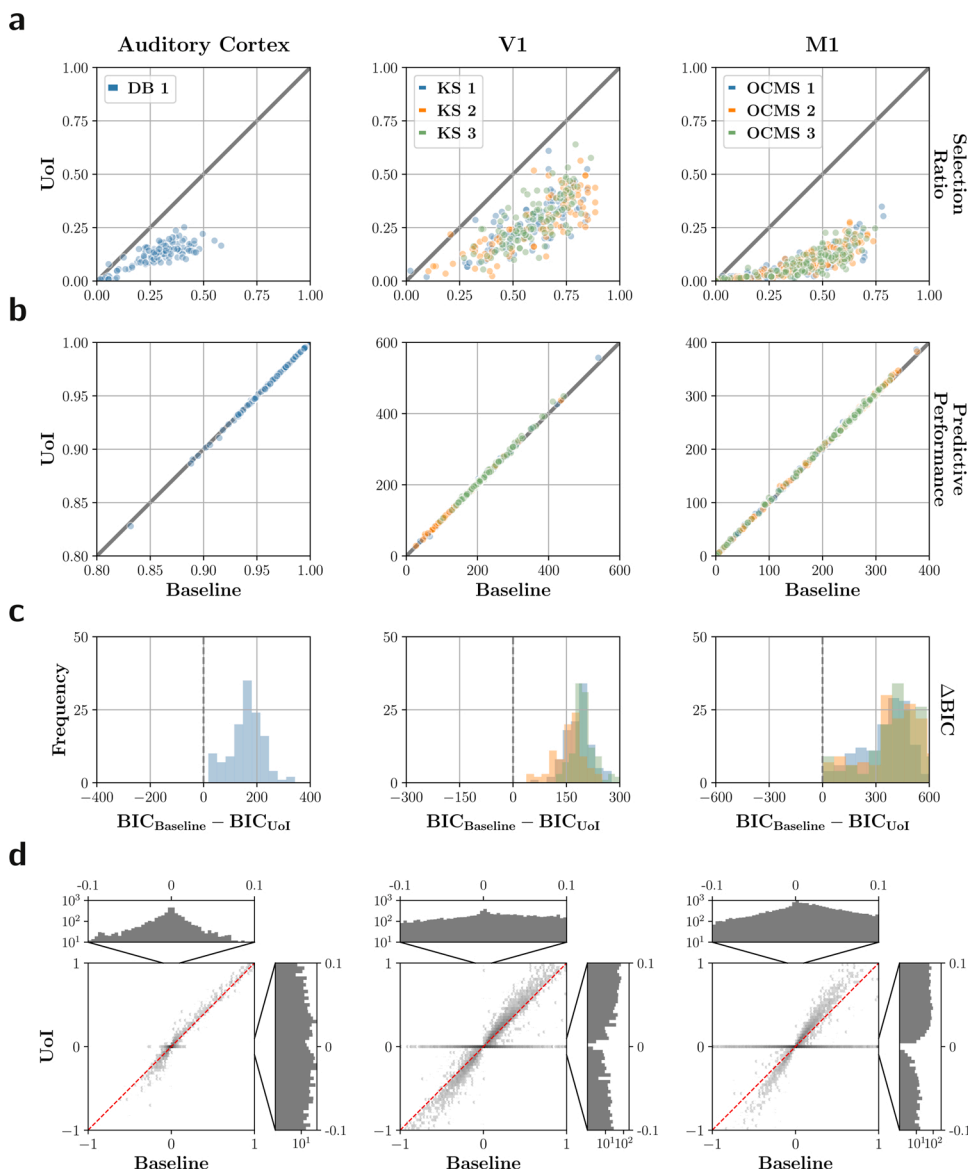
Fig. 4. Highly sparse coupling models maintain predictive performance. a-c Comparison of coupling models fit using baseline and UoI approaches to recordings from the auditory cortex (AC), primary visual cortex (V1), and primary motor cortex (M1). Each column corresponds to a different recording area (titles), and each row corresponds to a different model evaluation measure (right side labels). Colors denote different data sets within the same experiment, if available. Caption continued on following page. **Continued from previous page.** (a) The selection ratio, where the *y*-axis refers to the UoI model and *x*-axis the baseline model. Each point represents a coupling model for a specific single-unit/electrode. Gray line denotes the identity line. (b) Comparison of predictive performance. (c) The distribution of BIC differences across coupling models. (d) Comparison of the coefficient values inferred by UoI (*y*-axis) and baseline (*x*-axis) methods. Data is depicted on a hexagonal 2D histogram with a log intensity scale. The marginal distributions of the non-zero coefficient values are shown on the top (baseline) and side (UoI) for each brain area. Note that the 1D histograms display the distribution in a more restricted domain than the 2D histograms, as depicted by the black lines.

whether distributions of UoI and baseline model parameters were significantly different, we used the Kolmogorov-Smirnov test. We applied a significance level of $\alpha = 0.01$ for all statistical tests.

Cross-validation, model training, and model testing

Each dataset was split into 10 folds after shuffling across samples (except for the basal ganglia data, which was split into 5 folds due to fewer samples). When appropriate, the folds were stratified to contain equal proportions of samples across experimental setting (e.g., stimulus value or behavioral condition). In each task, we fit 10 models (or five, for basal ganglia) by training each on 9 (4) folds, and using the last fold as a test set. Hyperparameter selection for baseline procedures was performed via cross-validation within the training set of 9 (4) folds. Meanwhile, all resampling for the UoI procedures was also performed within the training set. Model evaluation statistics (selection ratio, predictive performance, Bayesian information criterion) are reported as the median across the 10 (5) models. Any measures that operate on the fitted models (e.g., coefficient value, network formation, modularity, etc.) were calculated by using the model that is formed by taking the median parameter value across folds.

3. Results

Parametric models are ubiquitous data analysis tools in systems neuroscience. However, their usefulness in understanding a neural system hinges on the assumption that their parameters are accurately selected and estimated. By accurate selection, we mean low false positives and false negatives in setting parameters equal to zero; by accurate estimation, we mean low-bias and low-variance in the parameter estimates. The potential neuroscientific consequences of improper selection or estimation during inference are generally not well understood. Thus, we studied selection and estimation in common systems neuroscience models by comparing the properties of models inferred by standard methods to those inferred by the Union of Intersections (UoI) framework. We fit models spanning functional coupling (coupling networks from auditory cortex, V1, and M1), sensory encoding (spatio-temporal receptive fields from retinal recordings and tuning curves from auditory cortex), and behavioral decoding (classifying behavioral condition from basal ganglia recordings). We analyzed the fitted models to assess whether improvements in inference impact the resulting neuroscientific conclusions.

3.1. Highly sparse coupling models maintain predictive performance

Functional coupling models detail the statistical interactions between the constituent units (e.g., neurons, electrodes, etc.) of a population. Such models can be used to construct networks, whose structural properties may elucidate the functional and anatomical organization of the neurons within the population (Song et al., 2013; Melozzi et al., 2019; Bassett and Sporns, 2017). Enhanced sparsity in these models could result in different inferred functional sub-networks reflected in the ensuing graph. Furthermore, obtaining biased parameter estimates obscures the relative importance of neuronal relationships in specific sub-populations. Therefore, precise selection and estimation in coupling models is necessary to properly relate the network structure to the statistical relationships between neurons.

We examined the possibility of building highly sparse and predictive coupling networks by fitting coupling models to data from three brain regions: recordings from auditory cortex (AC), primary visual cortex (V1), and primary motor cortex (M1). The AC data consisted of microelectrocorticography (μECoG) recordings from rat during the presentation of tone pips (Dougherty & Bouchard, 2019: DB). The V1 data consisted of single-unit recordings in macaque during the presentation of drifting gratings (Kohn & Smith, 2016: KS). The M1 data consisted of single-unit recordings in macaque during self-paced reaches on a grid of targets (O'Doherty, Cardoso, Makin, & Sabes: OCMS). See Methods for further details on experiments, model fitting, and metrics used for model evaluation, and see Table C.1 for a model statistic summary.

We constructed coupling models consisting of either a regularized linear model (AC) or Poisson model (V1, M1) in which the activity of an electrode/single-unit (i.e., node) was modeled using the activities of the remaining electrodes/single-units in the population. Thus, each dataset had as many models as there were distinct electrodes/single-units. We quantified the size of the fitted models with the selection ratio, or the fraction of parameters that were non-zero. We compare the selection ratio between baseline and UoI coupling models across electrodes/single-units in Fig. 4a. For all three brain regions, UoI models exhibited a marked reduction in the number of utilized electrodes/single-units. Specifically, UoI models used 2.24 (AC), 2.21 (V1), and 5.50 (M1) times fewer features than the corresponding baseline models. Across the populations of electrodes/neurons, this reduction was statistically significant ($p \ll 0.001$; see Table C.1) with large effect sizes (AC: $d = 1.74$; V1: $d = 2.26$; M1: $d = 2.49$). Interestingly, while the reduction in features for AC and V1 are roughly similar, the M1 models exhibit a much larger reduction in selection ratio, an observation that holds across the three M1 datasets. Furthermore, we examined coupling fits obtained via UoI$_{Lasso}$, finding that the enhanced sparsity persists despite the change in model. Additionally, the UoI linear and Poisson models exhibited similar recovery in selection profiles not recapitulated by baseline procedures (D).

We assessed whether the reduction in features resulted in meaningful loss of predictive accuracy. We measured predictive accuracy using the coefficient of determination (R^2) for linear models (AC) and the deviance for Poisson models (M1, V1), both evaluated on held out data. Note that in contrast to R^2, lower deviance is preferable. The predictive performances of baseline and UoI models for each brain region are compared in Fig. 4b. We observed that there is almost no change in the predictive performance across brain regions, with most points lying on or close to the identity line. We note that while the differences in performance across all models were statistically significant (AC: $p < 10^{-3}$; V1: $p \ll 0.001$; M1: $p \ll 0.001$; see Table C.1), the effect sizes of the reduction in predictive performance were very small (AC: $d = 0.005$; V1: $d = 0.05$; M1: $d = 0.03$), making it irrelevant in practice. Thus, these results imply that it is possible to construct highly sparse coupling methods that exhibit little to no loss in predictive performance across brain regions and datasets.

We captured the two previous observations — increased sparsity and

maintenance of predictive accuracy — with difference in Bayesian information criterion (BIC) between baseline and UoI methods, ΔBIC = BIC$_{baseline}$ − BIC$_{UoI}$. Lower BIC is preferable, so that positive ΔBIC indicates that UoI is the more parsimonious and preferred model. The distribution of ΔBIC across coupling models is depicted in Fig. 4c. ΔBIC is positive for all models, with a large median difference (AC: 170; V1: 149; M1: 186; see Table C.1). Thus, usage of BIC as a model selection criterion provides very strong evidence against the baseline models.

To characterize the functional relationships inferred by the coupling models, we examined the distribution of coefficient values. We normalized each model's coefficients by the coefficient with largest magnitude across the baseline and UoI models, and concatenated coefficients across models and datasets. We visualized the baseline and UoI coefficient values using a 2-d hexagonal histogram (Fig. 4d). First, we observed a density of bins above (positive coefficients) and below (negative coefficients) the identity line (Fig. 4d: red dashed line). This indicates that the magnitude of non-zero coefficients as fit by UoI are larger than the corresponding non-zero coefficient as fit by the baseline, demonstrating the amelioration of shrinkage and therefore reduction in bias. Next, we observed a density of bins on the $x = 0$ line, indicating a sizeable fraction of coefficients determined to be non-zero by baseline methods are set equal to zero by UoI. This density corroborates the reduction in selection ratio observed in Fig. 4a. We further note that the density of bins on the $x = 0$ line encompass a wide range of baseline coefficients values, especially for the V1 and M1 datasets. This implies that utilizing a thresholding scheme based on the magnitude of the fitted parameters for a feature selection procedure will not reproduce these results. Lastly, we observe no density of bins along the $y = 0$ line, which indicates that UoI models are likely not identifying the existence of functional relationships which do not exist (i.e., suffering from false positives).

While many of the coefficients set equal to zero by UoI have large magnitude (as measured by baseline methods), the bulk of density lies in coefficients with small magnitude. We found that the difference in distributions of non-zero coefficients between the two procedures is statistically significant ($p \ll 0.001$; Kolmogorov–Smirnov test). We highlight the marginal distribution of non-zero coefficients whose magnitudes are small (Fig. 4d, top and side histograms). While the baseline histograms (Fig. 4d, top histograms) have the largest density of coefficients close to zero, the UoI histograms, in a similar range, exhibit a large reduction in density. Together, these results demonstrate that coupling models fit by UoI possess qualitatively different parameter distributions, and raise the possibility that these differences may reflect shrinkage and abundance of false positives.

3.2. Improved inference enhances visualization, increases modularity, and decreases small-worldness in functional coupling networks

Functional coupling networks are useful in that they provide opportunities to visualize the statistical relationships within a population. Furthermore, their graph structures can be analyzed to characterize global properties of the network. The previous results show that improved inference gives rise to equally predictive models, but with much greater sparsity and qualitatively different parameter distributions. Thus, we next determined the impact on network visualization and structure. To this end, we constructed networks by placing coefficient values extracted from the coupling models directly in an adjacency matrix; i.e., $A_{ij} = \beta_{ij}$, where β_{ij} is the jth parameter for the ith coupling model. The adjacency matrices constituted directed graphs with weighted edges, which served as the primary focus in the subsequent analyses. When necessary, we considered undirected graphs calculated by symmetrizing pairwise coupling coefficients (see Methods for more details).

We first visualized the AC networks by plotting the baseline and UoI networks according to their spatial organization on the μECoG grid (Fig. 5a). Each vertex in Fig. 5a is color-coded by preferred frequency,
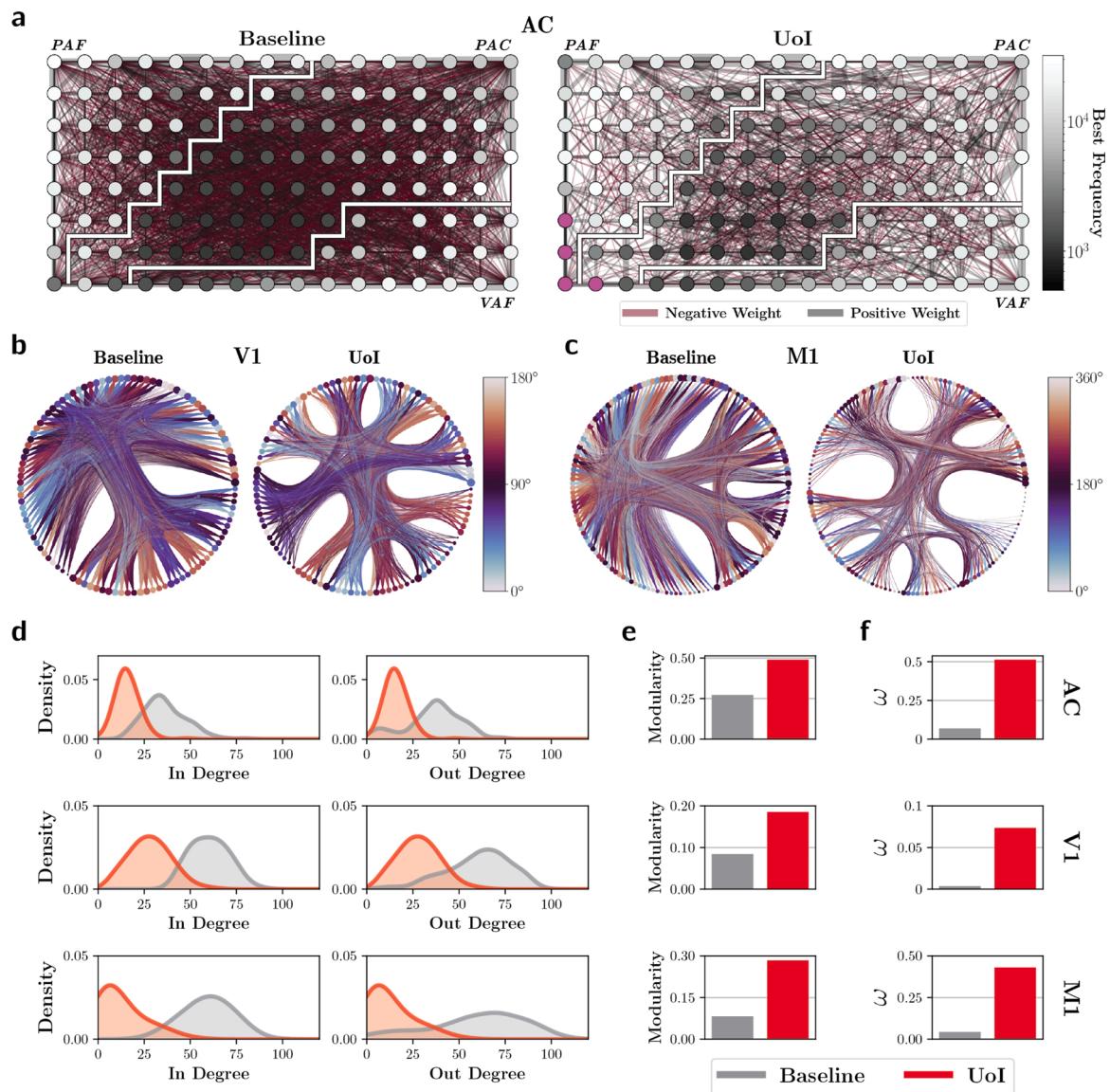
Fig. 5. Improved inference enhances visualization, increase modularity, and decrease small-worldness in functional coupling networks. (a) Networks obtained from auditory cortex data. Vertices are organized according to their position on the electrocorticography grid. Vertices are color coded by preferred frequency, with fuchsia vertices denoting non-tuned electrodes. Edge width increases monotonically with edge weight while edge color denotes the sign of the weight. White lines segment the grid according to the regions of auditory cortex. (b) Visualization of example coupling networks for visual cortex recordings, with vertices color-coded by preferred tuning. Edges are bundled according to detected communities, while vertex size corresponds to its degree. Note that stimulus encoding is cyclical (static grating angle from 0° to 180°). (c) Visualization of coupling networks for motor cortex recordings, with vertices color-coded by preferred tuning. Edges are bundled according to detected communities, and vertex size corresponds to its degree. Note that stimulus encoding is cyclical (movement direction angle from 0° to 360°). Caption continued on the following page. **Continued from previous page.** d-f. Comparison of common graph metrics evaluated on UoI networks (red) and baseline networks (gray). Each row corresponds to a distinct brain region (left: AC, V1, and M1 from top to bottom). (d) In-degree and out-degree densities. (e) The graph modularity, averaged across datasets. (f) The small-worldness as quantified by ω, and averaged across datasets.

while the symmetrized coupling coefficients are indicated by the color (sign) and weight (magnitude) of edges between vertices. We observed that the UoI network is easier to visualize, with densities of edges clearly demarcating regions of auditory cortex. This is contrast to the baseline network, whose lack of sparsity makes it difficult to extract any meaningful structure from the visualization. For example, the UoI network exhibits a clear increase in edge density in primary auditory cortex (PAC) relative to the posterior auditory field (PAF) and ventral auditory field (VAF). Thus, the increased sparsity in UoI networks reveals graph structure that ties in closely with general anatomical structure of the recorded region.

Since we lacked the physical layout for V1 and M1 recordings, we visualized the networks by fitting nested stochastic block models to the

directed graphs and plotting the ensuing structure in a circular layout with edge bundling (Fig. 5b, c). The nested stochastic block model identifies communities of vertices (neurons) in a hierarchical manner. We color-coded the vertices of the visualized graphs according to preferred tuning inferred from fitted encoding models (drifting grating angle for V1 networks and hand movement angle for M1 networks with cosine basis functions: see Methods). The UoI V1 network exhibits clear structure, with specific communities identifying similarly tuned neurons (Fig. 5b). The baseline V1 network does not exhibit as clear structure, and was highly unbalanced, with more than half the neurons placed in the same community. For the M1 data, the UoI communities were more balanced, though they lacked clear association with tuning properties. Together, these results demonstrate that enhanced sparsity of functional
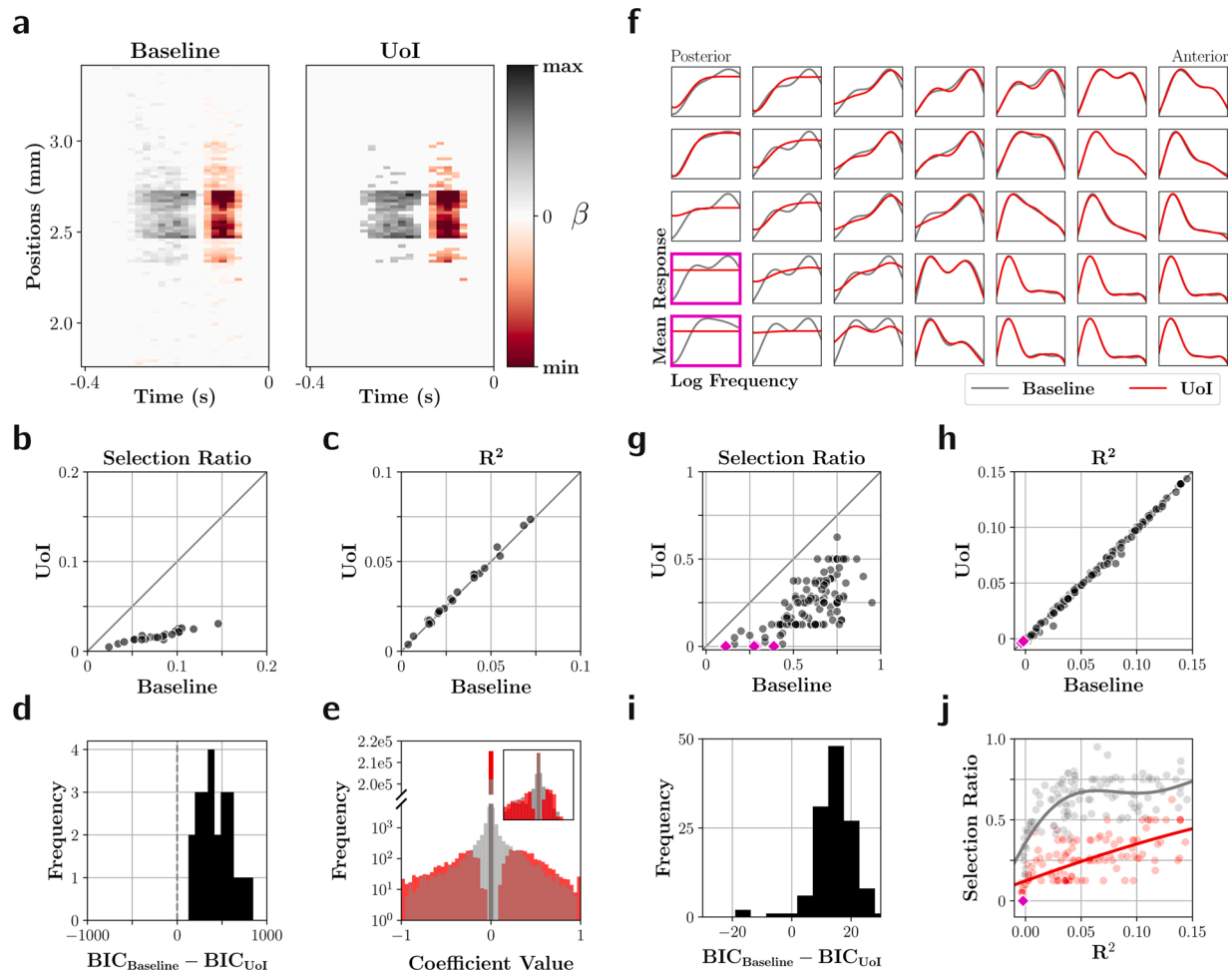
Fig. 6. Parsimonious tuning from encoding models. a-e Analysis of spatiotemporal receptive fields (STRFs) fit to spikes from mouse retinal ganglion cells using a 1d white noise stimulus. (a) Example STRF extracted by baseline (left) and UoI (right) procedures. (b-c) Quantitative comparison of STRFs extracted with UoI (y-axis) and baseline (x-axis) procedures. Each point represents a unique STRF. Gray line denotes identity. (b) Comparison of selection ratios for each STRF. (c) Comparison of coefficient of determination (R^2) on held out data. (d) The distribution in BIC differences across STRFs. Dashed line denotes equal BIC, i.e. ΔBIC $= 0$. (e) Distribution of normalized coefficient values across all baseline (gray) and UoI (red) STRFs. The broken y-axis is log-scale below the break and a regular scale above the break. Inset shows distribution of coefficients for STRFs depicted in **a. f-j.** Analysis of tuning curves extracted from micro-electrocorticography recordings on rat auditory cortex during the presentation of tone pips. (f) Examples of tuning curves for a subset of electrodes on the grid, as fit by baseline (gray) and UoI (red) procedures. Purple outlines denote tuning curves set exactly equal to zero by UoI (of the four electrodes, only two are shown). g-h. Quantitative comparison of tuning curves extracted with UoI (y-axis) and baseline (x-axis) procedures. Panels are structured similarly as panels **b-c.**. Purple points highlight tuning curves that had a selection ratio of zero, as determined by UoI. (i) The distribution in BIC differences across electrode tuning curves, structured similarly as panel (d) (j) Selection ratio plotted against coefficient of determination for baseline (gray) and UoI (red) procedures. Each point denotes a STRF. Trendlines are fit with Gaussian process regression.

coupling networks result in cleaner visualizations and, in the case of V1, a clearer connection to functional response properties.

These plots suggest different graph structures in the networks extracted by UoI and baseline methods. Thus, we first calculated the in-degree and out-degree distributions of the vertices in both networks (Fig. 5d). We observed that the in-degree and out-degree distributions for the UoI networks are much smaller, as one might expect due to the reduction in edges. Furthermore, the in- and out-degree distributions describing the UoI networks are nearly identical, in contrast to those of the baseline networks. Next, we calculated the modularity of the networks, which quantifies the degree to which the networks exhibit community-like structure. We found that the modularity for the UoI networks is much larger than that of baseline networks, indicating that UoI networks express more community structure than baseline networks (Fig. 5e). These results corroborate the visual findings in Fig. 5b. Since modularity implicitly depends on degree distribution, the enhanced community structure exhibited by the UoI networks is not simply a property of the reduction of in- and out-degrees, emphasizing that the enhanced sparsity is functionally meaningful. We found similar findings

in functional networks built from linear models, rather than Poisson models, implying that the structure of coupling networks persists across the type of underlying model (D).

Finally, we examined the small-worldness of the networks, a graph structure commonly used to describe brain networks. Small-world networks are characterized by a high degree of clustering and low characteristic path length, making them efficient for communication. We used ω to calculate small-worldness, whose values are bounded by the range $[-1, 1]$, with ω close to -1, 0, and 1 indicative of lattice, small-world, and random structure, respectively. Interestingly, UoI networks are considerably less small-world than the baseline networks (Fig. 5f). However, we note that all networks are more small-world than they are random. Furthermore, the small-worldness of the networks is dependent on the brain region. For example, the V1 networks exhibit substantially more small-worldness than the auditory cortex or M1 networks. Together, these results demonstrate that several properties of networks can be substantially altered by the utilized inference procedure, and that UoI networks are more modular and less small-world.

3.3. Parsimonious tuning from encoding models

A long-standing goal of neuroscience is to understand how the activity of individual neurons are modulated by factors in the external world (e.g., how the position of a moving bar is encoded by a neuron in the retina). In such encoding models, incorrect feature selection or parameter bias may mistakenly implicate factors in the production of neural activity, or misstate their relative importance. Thus, we examined how improved inference impacts tuning models, where an external stimulus is mapped to the corresponding evoked neural activity.

We first fit spatio-temporal receptive fields (STRFs) to single-unit recordings from isolated mouse retinal ganglion cells during the presentation of a flicker black or white bar stimulus (generated by a pseudorandom sequence). We used a linear model with a lasso penalty to fit STRFs (i.e., regularized, whitened spike-triggered averaging) to recordings from 23 different cells, using a time window of 400 ms. Thus, the fitted STRFs were two dimensional, with one dimensional capturing space (location in the bar stimulus) and the other capturing the time relative to neural spiking. For further experimental and model fitting details, see Methods. See Table C.2 for a dataset and model statistic summary.

The fitted STRFs for an example retinal ganglion cell are depicted in Fig. 6a. The UoI STRF captures the ON-OFF structure exhibited by the baseline STRF. However, the UoI model is noticeably sparser, resulting in a tighter spatial receptive field. The features set to zero by UoI (relative to baseline) include regions both further from the dominant ON-OFF structure, and regions very close to the center. Additionally, the coefficient values of the UoI STRF are noticeably larger in magnitude. These observations raise the possibility that the baseline procedure could be producing false positives in both the central and distal regions of the STRF, with the remaining coefficients suffering from shrinkage.

We compared the selection ratios across fitted STRFs (Fig. 6b) and found UoI fits to be substantially sparser, with a median reduction factor of 4.98. This reduction was statistically significant ($p \ll 0.001$; see Table C.2) and had a very large effect size ($d = 3.05$). At the same time, UoI models exhibited a statistically significant improvement in R^2 ($p < 0.01$; see Table C.2c), but with a very small effect size, making the improvement irrelevant in practice ($d = 0.05$; see C.2). Meanwhile, the BIC differences (Fig. 6d) were all large and positive (median difference $= 654$), providing very strong evidence in favor of the UoI model. Lastly, we compared the distribution of baseline and UoI encoding coefficients, normalized to the largest magnitude coefficient. We found evidence that the UoI models exhibited reduced shrinkage (Fig. 6e: larger tails), and a substantial reduction in false positives (Fig. 6e: reduced density at origin). Thus, improved inference resulted in STRFs with tighter structure.

Across neurons, we observed selection ratios and predictive performances spanning a wide range of values (Fig. 6b, c). We might expect that models with little predictive accuracy utilize fewer features, since poor predictive performance indicates that the provided features are inadequate. For example, in the limit that the model has no predictive capacity ($R^2 \leq 0$), the model should utilize no features, since such an R^2 indicates that none of the available features are relevant for reproducing the response statistics better than the mean response value. Therefore, we sought to determine whether inaccurate inference mistakenly identifies models as tuned (i.e., non-zero tuning features), when in fact a "non-tuned" model is appropriate (e.g., an intercept model: all features set equal to zero). To this end, we utilized a dataset in which the feature space dimensionality is small. This scenario provides a suitable test bed for assessing whether an intercept model could arise, and if such a model is appropriate given the response statistics of the data. We examined a dataset consisting of μECoG recordings from rat auditory cortex during the presentation of tone pips at various frequencies. We employed a linear tuning model mapping frequency to the peak (z-scored relative to baseline, see Methods), high-γ band analytic amplitude of each

electrode. The model features consisted of 8 Gaussian basis functions that tiled the log-frequency space.

We first examined whether improved inference resulted in any qualitative changes in the fitted encoding models. We plotted the fitted tuning curves as a function of log-frequency for a subset of electrodes arranged according to their location on the μECoG grid (Fig. 6f). Interestingly, the baseline and UoI tuning curves exhibit similar structure for a large fraction of the electrodes on the grid, in many cases matching closely (e.g., Fig. 6f: anterior side of grid). In other cases, particularly on the posterior side of the grid, the UoI tuning curves exhibit similar broad structure with noticeable smoothing, indicating that UoI has simplified the tuning model.

We compared the selection ratio of the models (Fig. 6g), and found that the UoI tuning curves utilize fewer features than those fit by baseline, with a median reduction factor of 2.5 that was statistically significant ($p \ll 0.001$; see Table C.2) and a large effect size ($d = 2.19$). Furthermore, despite a statistically significant decrease in R^2 (Fig. 6h) across electrodes ($p \ll 0.001$; see C.2), the observed effect size is very small (median $\Delta R^2 = 0.001$; $d = 0.05$). Meanwhile, we observed a median BIC difference of 19.4, providing evidence in favor of the UoI models (Fig. 6i). Taken together, these results imply that the reduction in features did not harm the predictive performance of the tuning models, thereby enhancing their parsimony.

We highlight four electrodes whose selection ratios, according to UoI, are exactly zero in Fig. 6g (purple points; some points overlap). The tuning curves of two of these electrodes are depicted in the posterior region of Fig. 6f (purple axis boundaries). These four "non-tuned" electrodes are among the least predictive, with R^2 close to or below zero for both baseline and UoI methods (Fig. 6h: purple points; some points overlap). Interestingly, the baseline selection ratio for one of these models was close to 0.4, indicating that UoI is generating models with substantially different support. We visually examined the frequency-response areas (FRAs) of these four electrodes, which detail the mean response values as a function of sound frequency and amplitude. We compared them to the FRAs of two randomly chosen electrodes, finding they had little discernible structure relative to the "tuned" FRAs (E). Thus, while increased sparsity in encoding models may not always result in perceptible changes in their appearance (e.g., Fig. 6f), there are cases where an inference procedure may mistakenly imply that a constituent unit is tuned, when in fact the stimulus features are not relevant for capturing its response statistics. To understand the behavior of the selection ratio as $R^2 \rightarrow 0$, we examined the relationship between selection ratio and R^2 for baseline (gray) and UoI (red) models (Fig. 6j). We found that the sparser models exhibit lower predictive power, with model trends predicting that at $R^2 = 0$, the selection ratio for the baseline model would be 0.35 ± 0.10 (mean ± 1 s.d.) while the UoI selection ratio would be 0.12 ± 0.10. This demonstrates that baseline procedures can suffer from false positives even when their fitted models exhibit little to no explanatory power. Overall, our results reveal that UoI can identify units as non-tuned when their encoding models lack predictive ability.

3.4. Decoding behavioral condition from neural activity with a small number of single-units

Decoding models describe which neuronal sub-populations contain information relevant for an external factor, such as a stimulus or a behavioral feature. Such models can identify which neurons may be useful to a downstream population for a task that requires knowledge of an external factor. Specifically, a decoding model's non-zero parameters can be interpreted as the sub-population of neurons containing the task-relevant information, emphasizing the need for precise selection. Additionally, the model details how specific neurons describe the decoded variable through the magnitudes of its parameters, requiring unbiased estimation. Thus, we sought to assess the degree to which an improved inference algorithm might impact data-driven discovery in
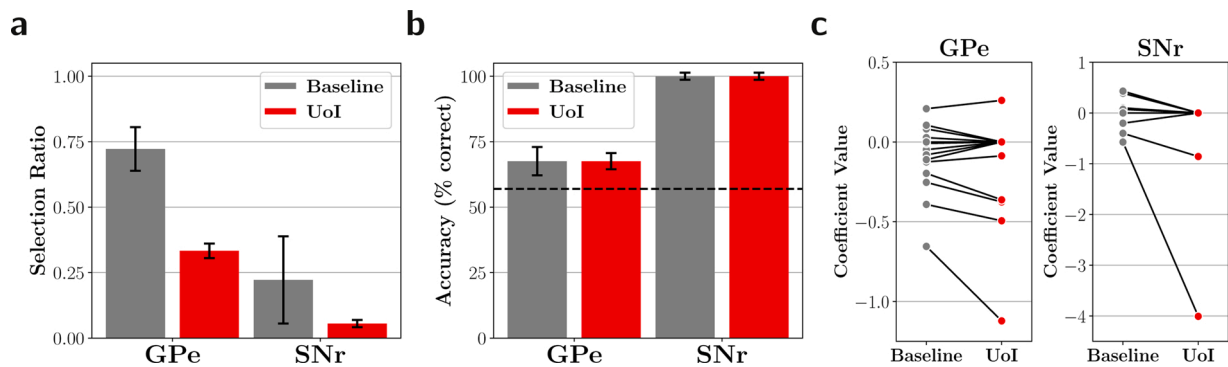
Fig. 7. Behavioral condition can be decoded with a small number of single-units at no loss in accuracy. Decoding models were applied to single-unit recordings from rat basal ganglia. The models, consisting of binary logistic regression, predicted whether the rat went left or right on a stop signal task. Single-unit recordings were used from the globus pallidus pars externa (GPe: 18 units) and the substantia nigra pars reticulata (SNr: 36 units). a-b. Evaluation of decoding models fit via the UoI (red) and baseline (gray) procedures. Bar heights indicate median across five data folds, while error bars denote IQR. (a) Comparison of the selection ratios. (b) Comparison of left/right classification accuracy on held-out data. Dashed line denotes accuracy by chance. (c) Comparison of fitted coefficient values extracted by baseline and UoI procedures in GPe and SNr. The decoding models fit by UoI utilize about 3 times fewer single-units at no cost to accuracy, indicating that task relevant information is contained in a small number of single-units.

neural decoding models.

For this analysis we examined 54 single units in the rat basal ganglia (18 units from globus pallidus pars externa, GPe, and 36 from the substantia nigra pars reticulata, SNr) that were recorded simultaneously during performance of a behavioral task involving rapid leftward or rightward head movements in response to cues. Details of the task are given in Gu et al. (2020); the analysis was restricted to trials in which a correct head movement was made. Thus, the decoding model consisted of binary logistic regression predicting trial outcome (left or right) using the single-unit spike counts as features. We fit the logistic regression with an ℓ_1 penalty (baseline) and the UoI framework (UoI$_{Logistic}$). Further details on the experimental setup and model fitting can be found in Methods. See C.3 for a summary of the dataset and fitted model statistics.

The selection ratios for GPe and SNr, as obtained by baseline and UoI procedures, are depicted in Fig. 7a. In GPe, the UoI decoding models utilized about half the number of parameters as the baseline procedures. Meanwhile, in SNr, the UoI models utilized four times fewer parameters than the baseline. Furthermore, we observe that UoI model sizes were more consistent across folds of the data. For example, the baseline SNr decoding models estimated anywhere from 1 to 21 parameters (out of 36) depending on the data fold, while UoI models consistently used only 2 or 3 parameters (Fig. 7a: IQR bars). These results validate that neural decoding models are capable of utilizing fewer features to predict relevant behavioral features. Furthermore, the stability principle ensures that these features are more robust to perturbations of the data (e.g., random subsamples).

To examine whether the use of fewer single-units decreased predictive performance, we evaluated the decoding models' classification accuracy on held-out data, depicted in Fig. 7b. The classification accuracy of the UoI models is equal to that of the baseline models for both GPe (67%) and SNr (100%). Furthermore, in both regions, the classification accuracy is greater than chance (56%), implying that the models are extracting meaningful information about the behavioral condition from the neural response. Interestingly, the median SNr models achieve perfect classification accuracy. The UoI model achieves this performance utilizing only 2 neurons, in contrast to median baseline model, which utilizes 8. Therefore, the activities of only a small subset of neurons are required to predict the behavioral condition, an observation that required an improved inference algorithm to consistently capture.

We examined the fitted coefficient values for each brain region and fitting procedure (Fig. 7c). First, we observed that all coefficients set equal to zero by the baseline procedure are also set equal to zero by UoI. Additionally, the coefficients set equal to zero by UoI, but not the

baseline procedure, typically have smaller magnitudes than the coefficients that are non-zero for both procedures. Finally, the coefficients set non-zero by UoI have larger magnitudes relative to their value under the baseline procedure. These observations imply that the UoI procedure, for this task, consistently utilized only the most important neurons to predict the behavioral condition. At the same time, UoI elevated the coefficient values relative to the baseline procedure, implying that baseline procedures suffered from substantial parameter shrinkage. In contrast to the coupling models, the parameters with the lowest magnitude in the baseline decoding models were set to zero by UoI. This could reflect differences in the task (classification vs. regression) or collinearity between the neural responses. Overall, these results demonstrate that task-relevant information is conveyed by a sparse subset of basal ganglia neurons, especially in SNr.

4. Discussion

Parametric models are used pervasively in systems neuroscience to characterize neural activity. The parameters of these models must be precisely selected and estimated in order to ensure accurate interpretation, particularly in the sparse parameter regime that is desirable for neural data. Motivated by the advent of new inference procedures capable of improved inference, we used the UoI framework to assess the degree to which poor parameter inference may impact neuroscientific interpretation of parametric models. We fit functional coupling, encoding, and decoding models to a battery of neural datasets, using standard and UoI inference procedures. We found that, across all models, the number of non-zero parameters could be reduced by a factor of 2–5, while maintaining predictive performance. Furthermore, we found broader, structural differences in the models beyond enhanced sparsity, which resulted in concrete changes to secondary analyses that inform neuroscientific interpretation.

The parameters obtained in coupling models denote the existence and strength of functional relationships between constituent units in a population. We fit coupling models that exhibited a marked reduction in model size, providing evidence that the baseline models suffered from false positives. Interestingly, the amount of feature reduction varied across brain areas, which may reflect differences in the nature of neural activity for these regions. Furthermore, we observed striking differences in the distribution of these parameter estimates, which impacted the graph structure. These changes produced cleaner visualizations, which in the case of V1, could related to the functional response properties of the neurons. In M1, we found no such relationship, inline with the dynamics view of this region (Churchland et al., 2012). Additionally, these

networks were characterized by increased modularity and decreased small-worldness. These results do not directly contradict previous work characterizing brain networks as small-world, but do reduce the magnitude of the characterization (Bassett and Bullmore, 2006, 2017; Honey et al., 2007) (though see Markov et al., 2013). Furthermore, coupling model parameters have been assessed for their recapitulation of synaptic weight distributions in neural circuits, in some cases identifying parameter biases induced by specific dynamical regimes of neural activity (Das and Fiete, 2019). These biases could be due to shrinkage of large parameter values, as observed here, and are mitigated by UoI. Furthermore, the coupling parameters extracted by UoI better reflect the weight distribution as observed in neural circuits, which is characterized by sparse connectivity with a heavy-tailed distribution (Song et al., 2005; Barbour et al., 2007). This was not achieved by baseline procedures, suggesting that the previously identified biases are due to inaccurate inference. The salient differences in the inferred coupling parameter distributions we observed motivates similar examination in models that capture neural dynamics, such as vector auto-regressive models (Balasubramanian et al., 2018; Ruiz et al., 2019), which could be further assessed by controllability metrics used in recent work on fMRI networks (Gu et al., 2015).

The parameters in encoding models detail which features modulate neural activity. We observed that the application of UoI to the encoding models highlighted cases where the fitted model had only zero parameter values, other than the intercept. Such an intercept model implies that a tuning model may be inappropriate for capturing the response statistics of the constituent units. This observation can be understood as a natural consequence of stability enforcement during parameter inference: UoI benefits from the stability principle by only utilizing selected features that persist across data resamples (Murdoch et al., 2019). The use of data resamples mimics perturbing the dataset, ensuring that features are included only if they are robust to those perturbations. Thus, the stability principle enforces model parsimony by encouraging the use of fewer features, eliminating those that offer no predictive accuracy throughout the resamples. However, UoI prioritizes predictive accuracy in the model averaging step. Therefore, models will only be made "as simple as possible" (e.g., removing all features) when they possesses no predictive ability. Furthermore, the fit spatio-temporal receptive fields on the retinal ganglion cells had typical ON-OFF structure, characteristic of the linear model (Pillow et al., 2008). In contrast to the auditory cortex, the stability enforcement of UoI resulted in models that were more spatially constrained, in better agreement with theoretical work characterizing the receptive fields and more accurately reflecting the visual features that explain the production of neural activity in retinal ganglion cells (Olshausen and Field, 1996; Stringer et al., 2019). More broadly, these results indicate that such improvements in parameter inference could serve to close the gap between experiment and theory in systems neuroscience.

Likewise, decoding models can inform which internal factors contain information about a task-relevant external factor. We found that decoding models could be fit using fewer single-units, at no cost to classification accuracy, implying that task-relevant information can be confined to a very small fraction of a neural circuit. In the context of this work, we note that SNr is a basal ganglia output nucleus, receiving converging inputs from multiple basal ganglia structures including GPe. Thus, the finding that SNr decodes behavioral output more selectively and accurately compared to GPe is consistent with the idea that SNr is closer to post-decision behavioral outputs, whereas GPe represent internal preparatory states. Our observations have more general implications for communication between brain areas: wiring constraints often restrict information transmission through a relatively smaller number of projection neurons, suggesting that these neurons contain the relevant information required to "decode" a given signal (Schüz et al., 2006; Isely et al., 2010). These results, in which we identified a very small fraction of the neurons capable of accurate decoding, raise the possibility that these selected neurons are, in fact, the projection neurons. Decoding

using fewer single-units also has practical implications. An abundance of work has considered the fidelity of a neural code by assessing the decoding ability of neural populations as a function of population size (Kohn et al., 2016). These decoding analyses can be informed by knowledge of the sparse sub-populations predictive of an external factor, which these results indicate are smaller than previously thought. Brain-machine interfaces (BMIs), which rely on accurate decoding from neural activity to operate, could reduce their power consumption by using a decoder relying on fewer single-units. Together, these results imply that accurate inference procedures, more capable of discovering specific task-relevant neuronal sub-populations, could drive the development of normative theories of neural communication and decoding.

Across brain regions and models, UoI resulted in more parsimonious models with differences in predictive performance that were irrelevant in practice, as measured by Cohen's *d*. However, statistical tests comparing the distribution of predictive performance between the baseline and UoI models revealed a statistically significant decrease in predictive performance for some cases (coupling models, AC tuning) and statistically significant increase in others (retinal STRF, decoding). Depending on one's goals, relying solely on predictive performance to judge a model may be unreliable (Shao, 1993; Wang et al., 2007). In particular, because model interpretability depends crucially on the included features and their estimates, we prioritized feature selection and estimation. Cross-validated predictive accuracy is often a poor criterion for accurate feature selection (Shao, 1997; Zhang et al., 2010). In these cases, the BIC, which captures model parsimony, serves as a more suitable criterion (Schwarz, 1978), and universally favored the UoI models (though we note there is no single preferred model selection criterion (Schwarz, 1978; Akaike, 1974; Rissanen, 1978; George and Foster, 2000; Zhang et al., 2010)). Furthermore, the increase in sparsity of the UoI models imply that the baseline fits suffered from an abundance of false positives. The similar predictive performance between the UoI and baseline fits could imply that false positives can have little impact on predictive performance in models of neural data, further emphasizing the need to consider multiple criteria in model selection. Alternatively, UoI fits may have suffered from false negatives, thereby lowering predictive performance, though previous experiments in synthetic data make this unlikely (Bouchard et al., 2017).

We considered models of neural activity exclusively in terms of coupling, encoding, or decoding. However, past studies have built parametric models of neural activity by using other features or model structures. For example, the combination of coupling and encoding in a single model is a natural extension which has been examined extensively in previous work (Stevenson et al., 2012; Pillow et al., 2008; Vidne et al., 2012; Truccolo et al., 2005; Okatan et al., 2005). Other features that are not constrained within coupling, encoding, or decoding — such as spike-time history or global fluctuations — are also important to incorporate (Truccolo et al., 2005; Okun et al., 2015; Paninski et al., 2007). Additionally, latent variable models have been used to great success to capture, in particular, the dynamics of neural activity (Byron et al., 2009; Paninski et al., 2010; Macke et al., 2011). In this work, the stability principles used by UoI resulted in a significant difference in the model sparsity and estimated parameter distribution, which impacted interpretation. Future work should assess whether similar results can be achieved in these extended models, and if so, determine the neuroscientific consequences.

UoI is formulated in a frequentist context, which guided the baseline methods we chose as comparisons. This was done to have maximal utilization by neuroscientists, which predominantly utilize frequentist inference. However, a multitude of inference approaches, including Bayesian methods (Chen et al., 2010; Zhou et al., 2016; Huang and Tran, 2018), have been introduced in recent years which all excel at parameter inference. This new class of inference approaches will fundamentally change the interpretation of parametric models on neural data, and therefore should preferentially be used in future studies to improve data-driven discovery. As these inference procedures continue to be

improved, additional assessments on neural data will be informative to better understand how neuroscientific interpretation further develops.

We restricted our analysis to generalized linear models, because their structure lends itself well to interpretation, making them ubiquitous in neuroscience. However, the improvements we obtained by encouraging stability and sparsity in these models may extend to other classes of models. For example, dimensionality reduction methods have also played an important role in systems neuroscience (Cunningham and Byron, 2014). The UoI framework is naturally extendable to such methods, including column subset selection (Bouchard et al., 2017) and non-negative matrix factorization (Ubaru et al., 2017). Furthermore, recent work has found success in using artificial neural networks (ANNs) to model neural activity, which excel at predictive performance (Kell and McDermott, 2019). Since ANNs are highly parameterized, these models are not interpretable in the sense that their parameter values do not directly convey neuroscientific meaning. Instead, these models are often interpreted through the lens of learned representations or recapitulation of emergent properties of neural activity. Future work could assess whether the inference principles described in this work could have similar effects for ANNs modeling neural activity, especially given recent advancements in compressing such models (Abbasi-Asl and Yu, 2017).

Competing interests

The authors declare no competing interests.

Author contributions

Pratik S. Sachdeva: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing

Jesse A. Livezey: Investigation, Software, Writing - review & editing
Maximilian E. Dougherty: Data Curation, Investigation
Bon-Mi Gu: Data Curation, Investigation, Writing - review & editing
Joshua D. Berke: Data Curation, Investigation, Resources, Supervision, Writing - review & editing

Kristofer E. Bouchard: Conceptualization, Funding acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing - original draft, Writing - review & editing

Appendix A Pseudocode for the Union of Intersections

We provide pseudocode to the Union of Intersections framework. This pseudocode is generalized to all UoI algorithms discussed in the paper, with the objective $L(\beta; X^k, y^k)$ corresponding to the choice of algorithm.

Algorithm 1.   Union of Intersections inference in generalized linear models

Input: Data $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$

1:	Regularization strengths $\{\lambda_j\}_{j=1}^q$
2:	Number of resamples N_S and N_E
3:	Loss function $L(\beta; X, y)$
4:	*Model Selection*
5:	**for** $k = 1$ to N_S **do**
6:	Generate resample X^k, y^k
7:	**for** $j = 1$ to q **do**
8:	$\widehat{\beta}^{jk} \longleftarrow$ Optimize $L(\beta; X^k, y^k) + \lambda_j|\beta|_1$ (ℓ_1-penalized objective)
9:	$S_j^k \longleftarrow \{i\}$ where $\widehat{\beta}_i^{jk} \neq 0$
10:	**end for**
11:	**end for**
12:	**for** $j = 1$ to q **do**
13:	$S_j \longleftarrow \bigcap_{k=1}^{N_S} S_j^k$ ▷ *Intersection*
14:	**end for**
15:	*Model Estimation*
16:	**for** $k = 1$ to N_E **do**
17:	Generate training (X_T^k, y_T^k) resample
18:	**for** $j = 1$ to q **do**
19:	$X_{T,j}^k \longleftarrow X_T^k$ with features S_j extracted.
20:	$\widehat{\beta}^{jk} \longleftarrow$ Optimize $L(\beta; X_{T,j}^k, y_T^k)$
21:	$\ell^{jk} \longleftarrow \text{BIC}(\widehat{\beta}^{jk}, X_{T,j}^k, y_T^k)$
22:	**end for**
23:	$\widehat{\beta}^k \longleftarrow \underset{\widehat{\beta}^{jk}}{\text{argmin}}\ \ell^{jk}$
24:	**end for**
25:	$\widehat{\beta}^* = \underset{k}{\text{median}}(\widehat{\beta}^k)$ ▷ *Union*
26:	**return** $\widehat{\beta}^*$

Appendix B. Extended synthetic results for Poisson and logistic variants

We conducted additional synthetic experiments evaluating the performance of UoI in the context of Poisson and logistic regression. As in the case of the linear model (detailed in Section 2.2), we generated data with $p = 300$ total parameters of which $k = 100$ were non-zero. The non-zero ground truth parameters were drawn from a parameter distribution characterized by exponentially increasing density as a function of parameter magnitude (Fig. 3b: gray histograms). We used $N = 1200$ samples generated according to the Poisson model and $N = 2400$ samples generated according to the logistic model (logistic regression typically requires more samples for convergence). We compared UoI against ℓ_1-penalized Poisson regression, fit by glmnet (Friedman et al., 2010), and ℓ_1-penalized logistic regression, fit by scikit-learn (Pedregosa et al., 2011). We did not conduct an analysis against a full battery of methods, as done in Section 2.2, because solvers beyond the ℓ_1-penalty are not readily available in the Poisson and logistic contexts.

We summarized our findings with similar metrics detailed in Section 2.2. Specifically, we quantified selection performance with the selection accuracy, bias with estimation error, variance with the estimation variability, predictive performance with the deviance and log-likelihood, and model parsimony with the Bayesian information criterion. Our results are shown in Fig. B.1, with boxplots capturing the distribution of metrics across 30 synthetic datasets (in the case of estimation variability, the distribution across parameters). We found that UoI exhibits increased selection accuracy (Fig. B.1a, b) and decreased bias (Fig. B.1c,d). At the same time, UoI exhibits comparable variability, with most parameters seeing an improvement, but a longer tail exhibiting increased variance (Fig. B.1e, f). These improvements in model sparsity and bias translated to improved prediction performance, as quantified by the deviance for Poisson regression (Fig. B.1g: lower is better) and the log-likelihood for logistic regression (Fig. B.1h: higher is better). Lastly, UoI exhibited markedly improved model parsimony as quantified by the Bayesian information criterion (Fig. B.1i: lower is better).
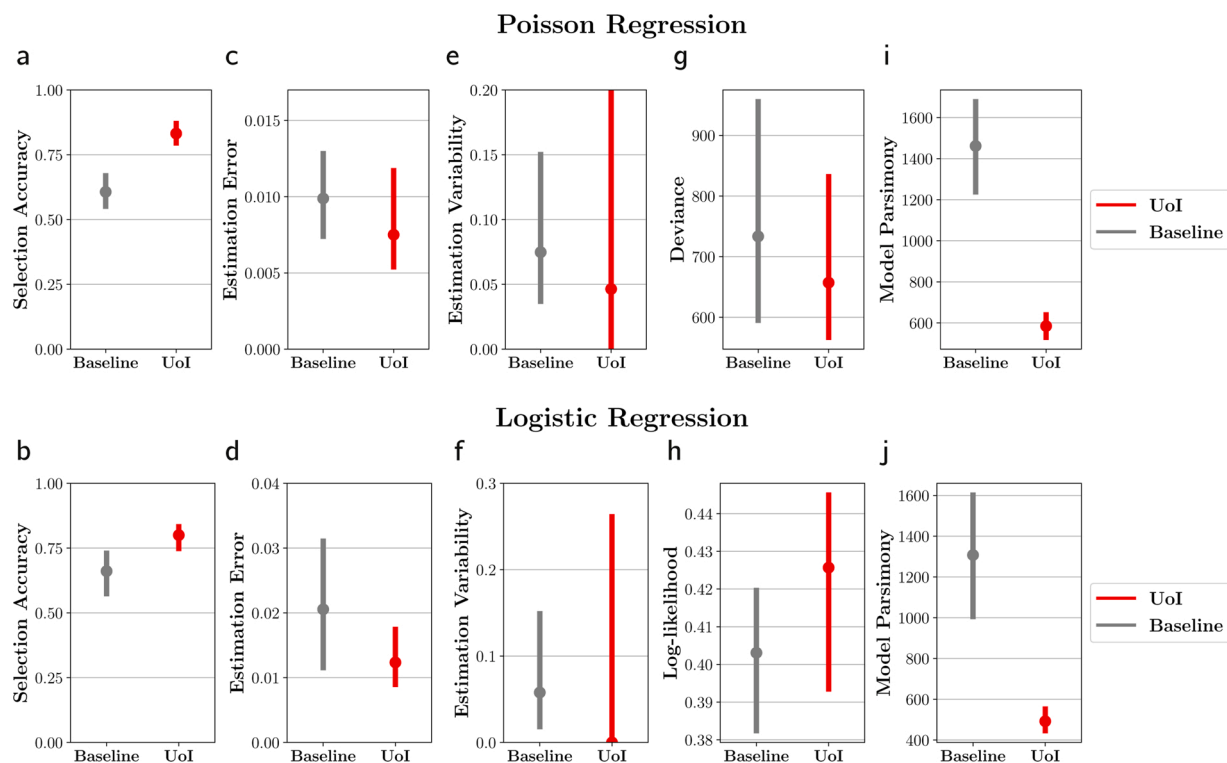

Fig. B.1. UoI exhibits improved selection, decreased bias, comparable variance, and superior model parsimony on Poisson and logistic variants. Top row depicts results for Poisson regression, while bottom row depicts results for logistic regression. Each column corresponds to a different metric, with red box plots denoting UoI performance and gray boxplots denoting baseline performance. Circular points denote median, with whiskers denoting the total spread of responses. a, b. Comparison of selection accuracies (higher is better). c, d. Comparison of estimation errors, quantifying bias (lower is better). e, f. Comparison of estimation variabilities (lower is better). g, h. Comparison of predictive performance. For Poisson models, this is quantified by deviance (g: lower is better). For logistic models, this is quantified by log-likelihood (h: higher is better). i, j. Comparison of model parsimonies, as quantified by Bayesian information criterion (lower is better).

Appendix C. Supplementary tables

C.1 Functional coupling dataset summary and fitted model statistics

Dataset summary (Table C.1a) provides details on the datasets used to fit functional coupling models, including the number of units and samples across brain region and recording session. The following tables provide statistics summarizing aspects of the fitted baseline and UoI models, including selection ratio (Table C.1b), predictive performance (Table C.1c), and Bayesian information criterion (Table C.1d).

C.2 Encoding model dataset summary and fitted model statistics.

Dataset summary (Table C.2a) provides details on the datasets used to fit encoding models, including the number of units and samples across dataset. The following tables provide statistics summarizing aspects of the fitted baseline and UoI models, including selection ratio (Table C.2b), predictive performance (Table C.2c), and Bayesian information criterion (Table C.2d).

C.3 Decoding model dataset summary and fitted model statistics

Dataset summary (Table C.3a) provides details on the datasets used to fit decoding models, including the number of units and samples across dataset. The following tables provide statistics summarizing aspects of the fitted baseline and UoI models, including selection ratio (Table C.3b), and predictive performance (Table C.3c).

Table C.1a
Dataset summary

Brain region	Number of units	Number of samples
AC	125	4200
VC 1	106	2400
VC 2	88	2400
VC 3	112	2400
MC 1	136	4089
MC 2	146	4767
MC 3	147	4400

Table C.1b
Selection ratio

Brain region	Baseline	UoI	Reduction factor	*p*-value	*d*
AC	0.30 ± 0.07	0.13 ± 0.03	2.24	5×10^{-23}	1.74
VC 1	0.59 ± 0.11	0.27 ± 0.07	2.21	2×10^{-19}	2.26
VC 2	0.66 ± 0.17	0.26 ± 0.11	2.56	2×10^{-16}	1.87
VC 3	0.59 ± 0.10	0.28 ± 0.07	2.13	2×10^{-20}	2.57
MC 1	0.43 ± 0.12	0.07 ± 0.04	5.85	3×10^{-24}	2.49
MC 2	0.46 ± 0.12	0.08 ± 0.04	5.50	3×10^{-25}	2.37
MC 3	0.46 ± 0.10	0.09 ± 0.04	5.15	1×10^{-25}	2.58

Table C.1c
Predictive performances

Brain region	Baseline	UoI	*p*-value	*d*
AC	0.98 ± 0.02	0.98 ± 0.02	5×10^{-4}	0.005
VC 1	203 ± 35	205 ± 36	3×10^{-16}	0.05
VC 2	166 ± 44	168 ± 43	3×10^{-12}	0.03
VC 3	213 ± 42	217 ± 41	2×10^{-18}	0.05
MC 1	232 ± 65	233 ± 65	1×10^{-23}	0.03
MC 2	263 ± 60	266 ± 59	1×10^{-23}	0.03
MC 3	248 ± 57	250 ± 58	2×10^{-25}	0.03

Table C.1d
Bayesian information criterion

Brain region	Baseline	UoI	Median difference
AC	− 6240 ± 1884	− 6491 ± 1859	170
VC 1	− 15349 ± 20946	− 15516 ± 20970	149
VC 2	− 2864 ± 11046	− 3007 ± 11087	131
VC 3	− 35164 ± 34069	− 35309 ± 34066	161
MC 1	481 ± 124	321 ± 59	162
MC 2	542 ± 150	354 ± 75	186
MC 3	564 ± 120	362 ± 67	190

Table C.2a
Dataset summary

Brain region	Number of units	Number of samples
Retina	125	4200
AC	23	89896

Table C.2b
Selection ratio

Brain region	Baseline	UoI	Reduction factor	p-value	d
Retina	0.084 ± 0.020	0.017 ± 0.004	4.98	3×10^{-5}	2.21
AC	0.625 ± 0.116	0.250 ± 0.114	2.50	7×10^{-23}	2.19

Table C.2c
Predictive performances

Brain region	Baseline	UoI	p-value	d
Retina	0.028 ± 0.013	0.028 ± 0.013	0.004	0.05
AC	0.042 ± 0.034	0.041 ± 0.033	2×10^{-19}	0.05

Table C.2d
Bayesian information criterion

Brain region	Baseline	UoI	Median difference
Retina	− 1646606 ± 44936	− 1647261 ± 44857	654
AC	4371 ± 2495	4351 ± 2493	19.4

Table C.3a
Dataset summary

Brain region	Number of units	Number of samples
GPe	18	186
SNr	36	186

Table C.3b
Selection ratio

Brain region	Baseline	UoI	Reduction factor
GPe	0.722 ± 0.083	0.333 ± 0.028	2.167
SNr	0.222 ± 0.167	0.056 ± 0.014	4

Table C.3c
Predictive performances.

Brain region	Baseline	UoI
GPe	0.676 ± 0.054	0.676 ± 0.031
SNr	1.000 ± 0.014	1.000 ± 0.014

Appendix D. Comparison of Poisson and linear coupling models for single-unit activity

We used a Poisson distribution to model single-unit spike count activity in the M1 and V1 datasets. However, past work has modeled single-unit activity with a linear-Gaussian model, after applying variance-stabilizing square root transform to the spike count responses. The degree to which a linear model can capture the functional relationships identified by a Poisson model for spiking data is unclear. Thus, we sought to characterize this capability, and its dependence on the inference procedure. We modeled the neural activity after a square-root transform $\sqrt{n_i}$ using a linear model

$$\sqrt{n_i} = \beta_0 + \sum_{j=1}^{M} \beta_{ij} \sqrt{n_j} + \varepsilon. \tag{D.1}$$

We fit this model with lasso optimization by coordinate descent (baseline) and UoI$_{\text{Lasso}}$.

We compared the fitted selection profiles, i.e. the set $S = \{i | \beta_i \neq 0\}$ between the linear and Poisson models. To do so, we used the hypergeometric distribution, which describes the probability that k objects with a particular feature are drawn from a population of size M that has K total objects with that feature, using m draws without replacement. To frame this in terms of selection, suppose the Poisson model has $|S_{\text{Poisson}}| = K$ non-zero parameters out of the M possible features. Then, the probability the linear model, which has $|S_{\text{linear}}| = m$ non-zero parameters, would match the Poisson model on k such features is given by the hypergeometric distribution:

$$\Pr(k) = \frac{\binom{K}{k}\binom{M-k}{m-k}}{\binom{M}{m}}. \tag{D.2}$$

Thus, the probability that the selection profile would overlap at most as well by chance as the linear model is given by $p_{\text{overlap}} = 1 - \Pr(k < k_{\text{linear}})$. We compared the distribution of p_{overlap} across coupling models, calculated for both baseline and UoI procedures (Fig. D.1, panel a). For the V1 data, the UoI linear models better reproduced the Poisson selection profiles, with 98% of the profiles fit by UoI satisfying $p_{\text{overlap}} < 0.001$, compared to only 74% of the baseline selection profiles. In contrast, in the M1 data, both inference procedures fit linear models that closely matched the selection profiles of the Poisson models, with 99% of the selection profiles satisfying $p_{\text{overlap}} < 0.001$. Therefore, improved inference results in more consistent selection across models, and furthermore this consistency depends on brain region.

We constructed networks from the linear models as described in Methods, and calculated their in-degree and out-degree distributions. We compared the in-degree and out-degree distributions of the linear networks to the Poisson networks, finding a closer correspondence between the UoI models than for the baseline models in most cases (Fig. D.1, panel b). Specifically, the in-degrees of UoI models had a correlation of 0.742 (V1) and 0.969 (M1) compared to 0.717 (V1) and 0.924 (M1) for the baseline procedures. Similarly, we obtained out-degree correlations of 0.806 (UoI) and 0.531 (baseline) for V1 and 0.918 (UoI) and 0.924 (baseline) for M1. Lastly, we found that the UoI linear networks were more modular than the baseline linear networks. Interestingly, both were more modular than their Poisson counterparts (Fig. D.1, panel c). Taken together, these results imply that a more precise inference framework better preserves structure across model types.
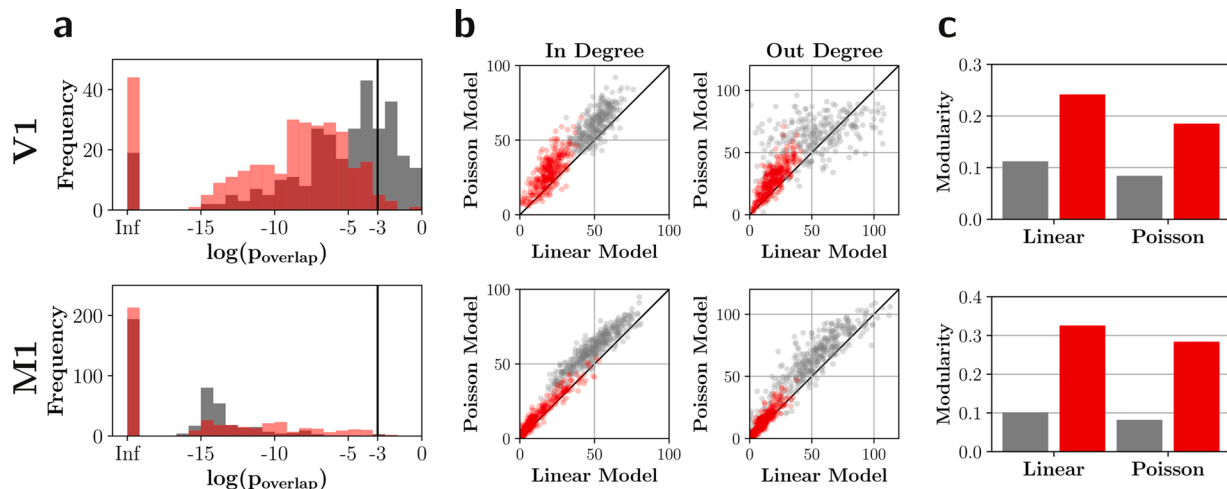

Fig. D.1. Improved inference ensures that the structure of fitted coupling networks persists across the type of underlying model. Linear and Poisson coupling models were fit to the datasets with single-unit recordings (V1 and M1) using baseline (gray) and UoI (red) procedures. Top row corresponds to results from networks fit to V1 recordings, while bottom row corresponds to networks fit to M1 recordings. (a) The (log) probability distribution of extracting a support (set of non-zero parameters) by the linear model that matches with the Poisson model support, according to a hypergeometric distribution. Vertical line denotes a p-value of 0.001 (b) Comparison of the in-degree and out-degree distributions between the Poisson network (y-axis) and linear network (x-axis). Each point represents a single unit. Black line denotes identity. (c) Graph modularity of linear and Poisson networks.

Appendix E. Frequency response area analysis for tuned and non-tuned electrodes
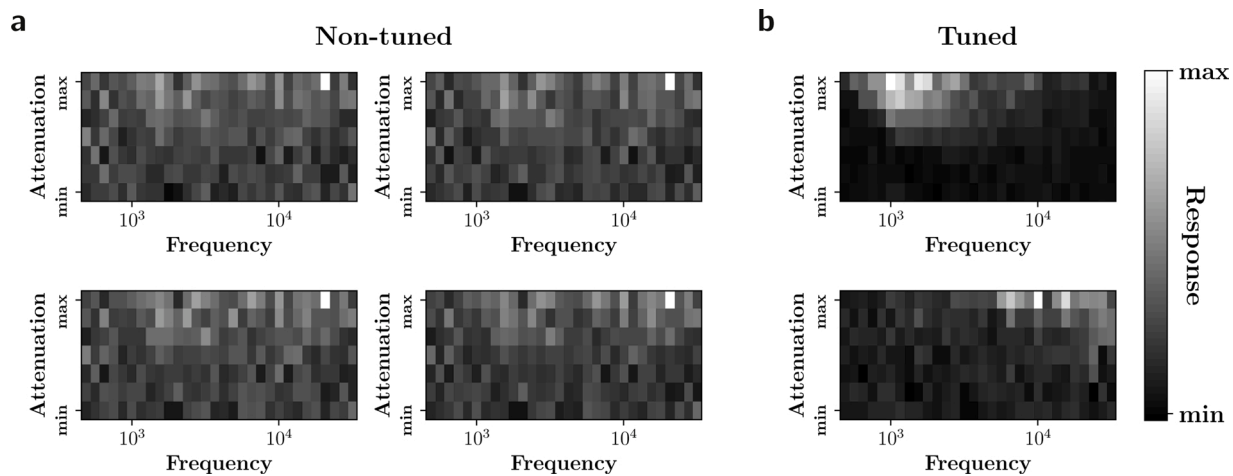
Fig. E.1

Fig. E.1. Frequency response area (FRA) analysis of non-tuned electrodes, as determined by UoI, confirm that a frequency tuning model captures no discernible structure in their responses. Each plot depicts the FRA, or the mean responses, across trials, to frequency-attenuation stimulus pairings. The plots are normalized to their maximum value. In each plot, the *y*-axis denote attenuation (ranging from −70 db to 0 db) while the *x*-axis denotes frequency (ranging from 500 Hz to 32 kHz). (a) FRAs corresponding to the four electrodes that UoI determined to be non-tuned (pink points in Fig. 6). (b) FRAs for two randomly chosen tuned electrodes. The non-tuned electrodes exhibit no discernible structure in the FRAs (in contrast to the clear structure depicted by the tuned electrodes), validating that UoI correctly determined them to be non-tuned.

References

Abbasi-Asl, R., Yu, B., 2017. Structural compression of convolutional neural networks based on greedy filter pruning arXiv:1705.07356.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control 19 (6), 716–723.

Anumanchipalli, G.K., Chartier, J., Chang, E.F., 2019. Speech synthesis from neural decoding of spoken sentences. Nature 568 (7753), 493–498.

Arnemann, K.L., Stöber, F., Narayan, S., Rabinovici, G.D., Jagust, W.J., 2018. Metabolic brain networks in aging and preclinical alzheimer's disease. NeuroImage: Clinical 17, 987–999.

Bühlmann, P., Van De Geer, S., 2011. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media.

Babadi, B., Casti, A., Xiao, Y., Kaplan, E., Paninski, L., 2010. A generalized linear model of the impact of direct and indirect inputs to the lateral geniculate nucleus. J. Vision 10 (10), 22–22.

Balasubramanian, M., Ruiz, T., Cook, B., Bhattacharyya, S., Shrivastava, A., Bouchard, K., et al., 2018. Optimizing the union of intersections lasso (uoi-lasso) and vector autoregressive (uoi-var) algorithms for improved statistical estimation at scale arXiv:1808.06992.

Baldassarre, L., Pontil, M., Mour ao-Miranda, J., 2017. Sparsity is better with stability: Combining accuracy and stability for model selection in brain decoding. Front. Neurosci. 11, 62.

Barabási, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286 (5439), 509–512.

Barbour, B., Brunel, N., Hakim, V., Nadal, J.-P., 2007. What can we learn from synaptic weight distributions? Trends Neurosci. 30 (12), 622–629.

Bassett, D.S., Bullmore, E., 2006. Small-world brain networks. The Neuroscientist 12 (6), 512–523.

Bassett, D.S., Bullmore, E.T., 2017. Small-world brain networks revisited. The Neuroscientist 23 (5), 499–516.

Bassett, D.S., Sporns, O., 2017. Network neuroscience. Nature Neurosci. 20 (3), 353.

Bouchard, K.E., Chang, E.F., 2014. Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography. in: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society;1; IEEE 6782–6785.

Bouchard, K.E., Mesgarani, N., Johnson, K., Chang, E.F., 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495 (7441), 327.

Bouchard, K., Bujan, A., Roosta, F., Ubaru, S., Prabhat, M., Snijders, A., Mao, J.-H., Chang, E., Mahoney, M.W., Bhattacharya, S., 2017. Union of intersections (uoi) for interpretable data driven discovery and prediction. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems;1; Vol. 30. Curran Associates;1; Inc.

Bouchard, K.E., 2015. Bootstrapped adaptive threshold selection for statistical model selection and estimation arXiv preprint arXiv:1505.03511.

Bousquet, O., Elisseeff, A., 2002. Stability and generalization. Journal of machine learning research 2 (Mar), 499–526.

Breiman, L., 1996. Bagging predictors. Machine learning 24 (2), 123–140.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G., 2013. API design for machine learning software: experiences from the scikit-learn project. in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning 108–122.

Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. 10 (3), 186–198.

Byron, M.Y., Cunningham, J.P., Santhanam, G., Ryu, S.I., Shenoy, K.V., Sahani, M., 2009. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. in: Advances in neural information processing systems 1881–1888.

Bzdok, D., Yeo, B.T., 2017. Inference in the age of big data: Future perspectives on neuroscience. Neuroimage 155, 549–564.

Carmena, J.M., Lebedev, M.A., Crist, R.E., O'Doherty, J.E., Santucci, D.M., Dimitrov, D. F., Patil, P.G., Henriquez, C.S., Nicolelis, M.A., 2003. Learning to control a brain-machine interface for reaching and grasping by primates. PLoS Biol. 1 (2).

Chen, Z., Putrino, D.F., Ghosh, S., Barbieri, R., Brown, E.N., 2010. Statistical inference for assessing functional connectivity of neuronal ensembles with sparse spiking data. IEEE Trans. Neural Systems Rehabilitation Eng. 19 (2), 121–135.

Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., Shenoy, K.V., 2012. Neural population dynamics during reaching. Nature 487 (7405), 51–56.

Clauset, A., Newman, M.E., Moore, C., 2004. Finding community structure in very large networks. Phys. Rev. E 70 (6), 066111.

Cohen, J., 2013. Statistical Power Analysis for the Behavioral Sciences. Academic press.

Cunningham, J.P., Byron, M.Y., 2014. Dimensionality reduction for large-scale neural recordings. Nat. Neurosci. 17 (11), 1500.

Das, A., Fiete, I.R., 2019. Systematic errors in connectivity inferred from activity in strongly coupled recurrent circuits bioRxiv.

Dayan, P., Abbott, L.F., 2001. Theoretical neuroscience: computational and mathematical modeling of neural systems. MIT press.

Dougherty, M.E., Nguyen, A.P.Q., Baratham, V.L., Bouchard, K.E., 2019. Laminar origin of evoked ecog high-gamma activity. In: in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). doi: 10.1109/EMBC.2019.8856786, pp. 4391–4394.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Statist. Assoc. 96 (456), 1348–1360.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Friston, K., Moran, R., Seth, A.K., 2013. Analysing connectivity with granger causality and dynamic causal modelling. Curr. Opin. Neurobiol. 23 (2), 172–178.

Fulcher, B.D., Fornito, A., 2016. A transcriptional signature of hub connectivity in the mouse connectome. Proc. Natl. Acad. Sci. 113 (5), 1435–1440.

George, E., Foster, D.P., 2000. Calibration and empirical bayes variable selection. Biometrika 87 (4), 731–747.

Glaser, J.I., Chowdhury, R.H., Perich, M.G., Miller, L.E., Kording, K.P., 2017. Machine learning for neural decoding arXiv preprint arXiv:1708.00909.

Gong, P., Ye, J., 2015. A modified orthant-wise limited memory quasi-newton method with convergence analysis. in: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 276–284.

Gu, S., Pasqualetti, F., Cieslak, M., Telesford, Q.K., Alfred, B.Y., Kahn, A.E., Medaglia, J. D., Vettel, J.M., Miller, M.B., Grafton, S.T., et al., 2015. Controllability of structural brain networks. Nat. Commun. 6 (1), 1–10.

Gu, B.-M., Schmidt, R., Berke, J.D., 2020. Globus pallidus dynamics reveal covert strategies for behavioral inhibition bioRxiv.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC.

Holdgraf, C.R., Rieger, J.W., Micheli, C., Martin, S., Knight, R.T., Theunissen, F.E., 2017. Encoding and decoding models in cognitive electrophysiology. Front. Syst. Neurosci. 11, 61.

Honey, C.J., Kötter, R., Breakspear, M., Sporns, O., 2007. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. Proc. Natl. Acad. Sci. 104 (24), 10240–10245.

Huang, S., Tran, T.D., 2018. Sparse signal recovery via generalized entropy functions minimization. IEEE Trans. Signal Process. 67 (5), 1322–1337.

Hubel, D.H., Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. 160 (1), 106–154.

Isely, G., Hillar, C., Sommer, F., 2010. Deciphering subsampled data: adaptive compressive sampling as a principle of brain communication. in: Advances in neural information processing systems 910–918.

Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res. 15 (1), 2869–2909.

Kass, R.E., Amari, S.-I., Arai, K., Brown, E.N., Diekman, C.O., Diesmann, M., Doiron, B., Eden, U.T., Fairhall, A.L., Fiddyment, G.M., Fukai, T., Grün, S., Harrison, M.T., Helias, M., Nakahara, H., Teramae, J.-n., Thomas, P.J., Reimers, M., Rodu, J., Rotstein, H.G., Shea-Brown, E., Shimazaki, H., Shinomoto, S., Yu, B.M., Kramer, M. A., 2018. Computational neuroscience: Mathematical and statistical perspectives. Annu. Rev. Stat. ts Appl. 5 (1), 183–214.

Kell, A.J., McDermott, J.H., 2019. Deep neural network models of sensory systems: windows onto the role of task constraints. Curr. Opin. Neurobiol. 55, 121–132.

Kelly, R.C., Smith, M.A., Kass, R.E., Lee, T.S., 2010. Local field potentials indicate network state and account for neuronal response variability. J. Comput. Neurosci. 29 (3), 567–579.

Kiani, R., Cueva, C.J., Reppas, J.B., Newsome, W.T., 2014. Dynamics of neural population responses in prefrontal cortex indicate changes of mind on single trials. Curr. Biol. 24 (13), 1542–1547.

Kohn, A., Smith, M.A., 2016. Utah array extracellular recordings of spontaneous and visually evoked activity from anesthetized macaque primary visual cortex (v1). URL https://doi.org/10.6080/K0NC5Z4X.

Kohn, A., Coen-Cagli, R., Kanitscheider, I., Pouget, A., 2016. Correlations and neuronal population information. Annu. Rev. Neurosci. 39, 237–256.

Lefebvre, J.L., Zhang, Y., Meister, M., Wang, X., Sanes, J.R., 2008. γ. Development 135 (24), 4141–4151.

Lim, C., Yu, B., 2016. Estimation stability with cross-validation (escv). J. Comput. Graphical Stat. 25 (2), 464–492.

Liu, H., Roeder, K., Wasserman, L., 2010. Stability approach to regularization selection (stars) for high dimensional graphical models. In: in: 24th Annual Conference on Neural Information Processing Systems 2010. NIPS 2010.

Livezey, J.A., Bouchard, K.E., Chang, E.F., 2019. Deep learning as a tool for neural data analysis: speech classification and cross-frequency coupling in human sensorimotor cortex. PLoS Comput. Biol. 15 (9), e1007091.

Macke, J.H., Buesing, L., Cunningham, J.P., Byron, M.Y., Shenoy, K.V., Sahani, M., 2011. Empirical models of spiking in neural populations. in: Advances in neural information processing systems 1350–1358.

Makin, J.G., O'Doherty, J.E., Cardoso, M.M.B., Sabes, P.N., 2018. Superior arm-movement decoding from cortex with a new, unsupervised-learning algorithm. J. Neural Eng. 15 (2), 026010. https://doi.org/10.1088/1741-2552/aa9e95.

Markov, N.T., Ercsey-Ravasz, M., Van Essen, D.C., Knoblauch, K., Toroczkai, Z., Kennedy, H., 2013. Cortical high-density counterstream architectures. Science 342 (6158), 1238406.

Marx, V., 2013. Biology: The big challenges of big data.

Melozzi, F., Bergmann, E., Harris, J.A., Kahn, I., Jirsa, V., Bernard, C., 2019. Individual structural features constrain the mouse functional connectome. Proc. Natl. Acad. Sci. 116 (52), 26961–26969.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. 116 (44), 22071–22080. https://doi.org/10.1073/pnas.1900654116.

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

Neath, A.A., Cavanaugh, J.E., 2012a. The bayesian information criterion: background, derivation, and applications. WIREs Computational Statistics 4 (2), 199–203 arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.199, doi:https://doi.org/10.1002/wics.199. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.199.

Neath, A.A., Cavanaugh, J.E., 2012b. The bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics 4 (2), 199–203.

Nelder, J.A., Wedderburn, R.W., 1972. Generalized linear models. J. Royal Stat. Soc.: Series A (General) 135 (3), 370–384.

Newman, M.E., 2006. Modularity and community structure in networks. Proc. Natl. Acad. Sci. 103 (23), 8577–8582.

O'Doherty, J.E., Cardoso, M.M.B., Makin, J.G., Sabes, P.N., 2017. Nonhuman Primate Reaching with Multichannel Sensorimotor Cortex Electrophysiology (May 2017) doi: 10.5281/zenodo.583331. URL https://doi.org/10.5281/zenodo.583331.

Okatan, M., Wilson, M.A., Brown, E.N., 2005. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. Neural Comput. 17 (9), 1927–1961.

Okun, M., Steinmetz, N.A., Cossell, L., Iacaruso, M.F., Ko, H., Barthó, P., Moore, T., Hofer, S.B., Mrsic-Flogel, T.D., Carandini, M., Harris, K.D., 2015. Diverse coupling of neurons to populations in sensory cortex. Nature 521 (7553), 511–515. https://doi.org/10.1038/nature14273.

Olshausen, B.A., Field, D.J., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381 (6583), 607–609.

Pandarinath, C., O'Shea, D.J., Collins, J., Jozefowicz, R., Stavisky, S.D., Kao, J.C., Trautmann, E.M., Kaufman, M.T., Ryu, S.I., Hochberg, L.R., et al., 2018. Inferring single-trial neural population dynamics using sequential auto-encoders. Nature methods 15 (10), 805–815.

Paninski, L., Pillow, J., Lewi, J., 2007. Statistical models for neural encoding, decoding, and optimal stimulus design. In: in: Computational Neuroscience: Theoretical Insights into Brain Function. Vol. 165 of Progress in Brain Research, Elsevier, pp. 493–507.

Paninski, L., Ahmadian, Y., Ferreira, D.G., Koyama, S., Rad, K.R., Vidne, M., Vogelstein, J., Wu, W., 2010. A new look at state-space models for neural data. J. Comput. Neurosci. 29 (1-2), 107–126.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pillow, J.W., Shlens, J., Paninski, L., Sher, A., Litke, A.M., Chichilnisky, E., Simoncelli, E. P., 2008. Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature 454 (7207), 995.

Rao, R.P., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neurosci. 2 (1), 79–87.

Rich, E.L., Wallis, J.D., 2016. Decoding subjective decisions from orbitofrontal cortex. Nature Neurosci. 19 (7), 973–980 doi:10.1038/nn.4320. URL https://doi.org/10.1038/nn.4320.

Rissanen, J., 1978. Modeling by shortest data description. Automatica 14 (5), 465–471.

Ruiz, T., Balasubramanian, M., Bouchard, K.E., Bhattacharyya, S., 2019. Sparse, low-bias, and scalable estimation of high dimensional vector autoregressive models via union of intersections arXiv:1908.11464.

Sachdeva, P., Livezey, J., Tritt, A., Bouchard, K., 2019. Pyuoi: The union of intersections framework in python. J. Open Source Softw. 4 (44), 1799.

Sawilowsky, S.S., 2009. New effect size rules of thumb. J. Modern Appl. Stat. Methods 8 (2), 26.

Schüz, A., Chaimow, D., Liewald, D., Dortenman, M., 2006. Quantitative aspects of corticocortical connections: a tracer study in the mouse. Cerebral cortex 16 (10), 1474–1486.

Schwartz, O., Pillow, J.W., Rust, N.C., Simoncelli, E.P., 2006. Spike-triggered neural characterization. J. Vision 6 (4), 13–13.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.

Sejnowski, T.J., Churchland, P.S., Movshon, J.A., 2014. Putting big data to good use in neuroscience. Nat. Neurosci. 17 (11), 1440.

Seth, A.K., Barrett, A.B., Barnett, L., 2015. Granger causality analysis in neuroscience and neuroimaging. J. Neurosci. 35 (8), 3293–3297.

Shao, J., 1993. Linear model selection by cross-validation. J. Am. Stat. Assoc. 88 (422), 486–494.

Shao, J., 1997. An asymptotic theory for linear model selection. Statistica sinica 221–242.

Sharpee, T.O., 2013. Computational identification of receptive fields. Annu. Rev. Neurosci. 36 (1), 103–120.

Smith, M.A., Kohn, A., 2008. Spatial and temporal scales of neuronal correlation in primary visual cortex. J. Neurosci. 28 (48), 12591–12603.

Song, S., Sjöström, P.J., Reigl, M., Nelson, S., Chklovskii, D.B., 2005. Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS Biol. 3 (3).

Song, D., Wang, H., Tu, C.Y., Marmarelis, V.Z., Hampson, R.E., Deadwyler, S.A., Berger, T.W., 2013. Identification of sparse neural functional connectivity using penalized likelihood estimation and basis functions. J. Comput. Neurosci. 35 (3), 335–357.

Stevenson, I.H., Kording, K.P., 2011. How advances in neural recording affect data analysis. Nature Neurosci. 14 (2), 139.

Stevenson, I.H., Rebesco, J.M., Miller, L.E., Körding, K.P., 2008. Inferring functional connections between neurons. Curr. Opin. Neurobiol. 18 (6), 582–588.

Stevenson, I.H., Cherian, A., London, B.M., Sachs, N.A., Lindberg, E., Reimer, J., Slutzky, M.W., Hatsopoulos, N.G., Miller, L.E., Kording, K.P., 2011. Statistical assessment of the stability of neural movement representations. J. Neurophysiol. 106 (2), 764–774.

Stevenson, I.H., London, B.M., Oby, E.R., Sachs, N.A., Reimer, J., Englitz, B., David, S.V., Shamma, S.A., Blanche, T.J., Mizuseki, K., Zandvakili, A., Hatsopoulos, N.G., Miller, L.E., Kording, K.P., 2012. Functional Connectivity and Tuning Curves in Populations of Simultaneously Recorded Neurons. PLoS Comput. Biol. 8 (11) https://doi.org/10.1371/journal.pcbi.1002775.

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., Harris, K.D., 2019. High-dimensional geometry of population responses in visual cortex. Nature 571 (7765), 361–365.

Teeters, J.L., Harris, K.D., Millman, K.J., Olshausen, B.A., Sommer, F.T., 2008. Data sharing for computational neuroscience. Neuroinformatics 6 (1), 47–55.

Telesford, Q.K., Joyce, K.E., Hayasaka, S., Burdette, J.H., Laurienti, P.J., 2011. The ubiquity of small-world networks. Brain connectivity 1 (5), 367–375.

Theunissen, F., David, S., Singh, N., Hsu, A., Vinje, W., Gallant, J., 2001. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network: Comput. Neural Syst. 12 (3), 289–316.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Royal Stat. Soc.: Series B (Methodological) 58 (1), 267–288.

Triplett, M.A., Goodhill, G.J., 2019. Probabilistic encoding models for multivariate neural data. Front. Neural Circuits 13, 1.

Truccolo, W., Eden, U.T., Fellows, M.R., Donoghue, J.P., Brown, E.N., 2005. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. J. Neurophysiol. 93 (2), 1074–1089.

Ubaru, S., Wu, K., Bouchard, K.E., 2017. Uoi-nmf cluster: A robust nonnegative matrix factorization algorithm for improved parts-based decomposition and reconstruction of noisy data. In: in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). doi:10.1109/ICMLA.2017.0-152, pp. 241–248.

Van de Geer, S.A., et al., 2008. High-dimensional generalized linear models and the lasso. Ann. Stat. 36 (2), 614–645.

Vidne, M., Ahmadian, Y., Shlens, J., Pillow, J.W., Kulkarni, J., Litke, A.M., Chichilnisky, E., Simoncelli, E., Paninski, L., 2012. Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. J. Comput. Neurosci. 33 (1), 97–121.

Vinje, W.E., Gallant, J.L., 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287 (5456), 1273–1276.

Wander, J.D., Blakely, T., Miller, K.J., Weaver, K.E., Johnson, L.A., Olson, J.D., Fetz, E.E., Rao, R.P., Ojemann, J.G., 2013. Distributed cortical adaptation during learning of a brain-computer interface task. Proc. Natl. Acad. Sci. 110 (26), 10818–10823.

Wang, H., Li, R., Tsai, C.-L., 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94 (3), 553–568.

Wang, H.E., Bénar, C.G., Quilichini, P.P., Friston, K.J., Jirsa, V.K., Bernard, C., 2014. A systematic framework for functional connectivity measures. Front. Neurosci. 8, 405.

Wasserman, L., Roeder, K., 2009. High dimensional variable selection. Ann. Stat. 37 (5A), 2178.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393 (6684), 440.

Wilcoxon, F., 1992. Individual comparisons by ranking methods. in: Breakthroughs in statistics 196–202.

Yamashita, O., Sato, M.-a., Yoshioka, T., Tong, F., Kamitani, Y., 2008. Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns. NeuroImage 42 (4), 1414–1429.

Yu, S., Huang, D., Singer, W., Nikolić, D., 2008. A small world of neuronal synchrony. Cerebral Cortex 18 (12), 2891–2901.

Yu, B., et al., 2013. Stability. Bernoulli 19 (4), 1484–1500.

Zhang, Y., Li, R., Tsai, C.-L., 2010. Regularization parameter selections via generalized information criterion. J. Am. Stat. Assoc. 105 (489), 312–323.

Zhang, Y.-F., Asari, H., Meister, M., 2014. Multi-electrode recordings from retinal ganglion cells. URL https://doi.org/10.6080/K0RF5RZT.

Zhao, M., Batista, A., Cunningham, J.P., Chestek, C., Rivera-Alvidrez, Z., Kalmar, R., Ryu, S., Shenoy, K., Iyengar, S., 2012. An l 1-regularized logistic model for detecting short-term neuronal interactions. J. Comput. Neurosci. 32 (3), 479–497.

Zhou, B., Moorman, D.E., Behseta, S., Ombao, H., Shahbaba, B., 2016. A dynamic bayesian model for characterizing cross-neuronal interactions during decision-making. J. Am. Stat. Assoc. 111 (514), 459–471.

Zhu, M., Rozell, C.J., 2013. Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system. PLoS Comput. Biol. 9 (8).