We thank the reviewers for their thoughtful and detailed comments. Brief replies follow below.

R1) Input dimensionality: P-GAMs have similar scaling with the number of input dimensions as traditional GLMs. Nonetheless, using P-GAMs makes most sense when the inputs are task or cognitive variables; we don't envision our methods being commonly used for estimating V1 RFs. This does not mean that P-GAMs are not generally useful: many of the new datasets in systems neuroscience (higher cortices, hippocampus) naturally lend themselves to our analysis.

Spatio-temporal filters: We chose to include temporal filters only for binary events, which helps interpretability in the context of our task. P-GAMs can also model spatio-temporal effects: we simply need to add one extra temporal filter for each input dimension, somewhat similar in flavor to the factorization in [Park & Pillow, 2013]. Alternatively, we can define 2D nonlinear spatio-temporal filters for each input dimension, each contributing one additive term to the final GAM expression. Both versions scale linearly with the number of inputs (with a potentially large hidden constant in the second version). We will include examples of this functionality in the new version of the paper (results and code demo).

Smoothness priors: our current regularizer already encourages smoothness. Incorporating general GP priors seems more difficult. GP regression is known to scale unfavorably with the number of inputs, requiring additional structure (e.g. Kronecker) to keep computation tractable [Wilson et al, 2014]; even with the extra tricks, GP-based tuning estimates are restricted to low dimensional inputs (less than 10, inapplicable to our data, see e.g. [Savin & Tkacik 2016]). A new variant of GP-based GAMs may prove competitive [Adam, Durrande & John, 2018] but, to our knowledge, this has not yet been adapted to neural tuning estimation. We will cover the GP literature in the discussion.

Fig1E: shows misses and false positives for detecting whether an input dimension drives neural responses or not.

Fig3C: We've done additional statistics to disentangle the role of mean firing on tuning and coupling strength. A partial correlation analysis confirmed that coupling patterns cannot be trivially explained by differences in mean firing rates.

R2) Neural implications: We could not elaborate on monkey data results due to space limits; journal paper to follow.

GLM comparison: We chose to compare P-GAMs against a vanilla 'Pillow GLM' instead of a fancier version because this is what most neuroscientists end up using in practice; the group sparsity regularization would be closest in aims, but, to our knowledge, that has not been extensively applied to real data. We are happy to expand on the links to these alternative methods in the introduction/discussion.

R3) Usability, adapting the code to new datasets: Up to now, we have used the estimator on 3 different datasets (the monkey one, rat hippocampus and OFC). The initial P-GAM model specification takes a bit of work on any new dataset, but there are relatively few knobs that have to be set by hand (the spline basis); once the model is set, the code runs smoothly — a rotation student with limited coding background managed to get everything done in a few weeks.

ARD: The ARD regularizer corresponds to a factorized Gaussian prior with β-specific variances (treated as hyperparameters). Assuming a 1-to-1 map between parameters and inputs, the prior for irrelevant input dimensions will end up sharply concentrated around zero. It is not clear how to do this in the nonlinear case, when several βs are used to model each input dimension (one per basis vector). We need some form of group sparsity and traditional ARD can't do that.

R4) GLM vs P-GAM on real data: we can include GLM filter estimates and additional P-GAM vs. GLM fit quality quantification on real data in the supplementary info. Full GLM model comparison is intractable for this dataset so we can only do elastic net regularization. In brief, everything is much messier, although some of the trends persist.

Validation of CIs: Fair enough. We can definitely get bootstrapping-based CIs for artificial data, probably also for real units although it may prove too computationally expensive to do extensively. We will add those in the updated version. Apart from that, please note that the numerical estimates for the type 1 and 2 errors made by our significance test (Fig3C) are sensible – even if, admittedly, we could only do the validation for artificial data.

Utility and relevance: In practice, there are many experimental scenarios where simple GLMs don't quite suffice; most experimentalists will find the technical details too intimidating to attempt complex hierarchical regularization; they also often don't have the computational resources to do brute force model comparison for large models. We are providing a straightforward and relatively general way to get the job done, one that stays tractable even for large datasets.