To all reviewers. We thank the reviewers for the positive evaluation of our work and for all the suggestions for improving the clarity of our manuscript. Firstly, let us first state clearly that, upon acceptance, we will happily share our code and data. Many reviewers' concerns focused on the correction procedure of Eq. 4. This correction is necessary because when we learn the stimulus filters (step-1) the couplings are set to zero. However, when we include them in the *full model* step, they generate a mean contribution to the PSTH that needs to be subtracted. Otherwise this contribution would sum up to the stimulus signal and cause an over-estimation of the PSTH. More intuitively, and as reviewer 3 points out, the internal dynamics and so the coupling network can also drive the system, and in the step-1 of our procedure we neglect this. This is exactly why we need to introduce the correction of Eq. 4, which accounts for the contribution of the network to the mean response (PSTH). In addition, the correction term of Eq. 4 can be seen as a first order approximation of an optimal Plefka expansion (Plefka 1982, Kappen and Rodirguez 1997) around vanishing couplings. It could be possible to extend our method to second (called TAP) or higher orders. However, as we did not observe significant improvements on our results, we decided to keep the first order only, and to present it as a simpler correction accounting for the mean effect. Note that Eq. 4 just subtracts the mean of the left hand side. We will take care of explaining all this better.

Reviewer 1. Thanks, we have nothing to add.

Reviewer 2. The reviewer's major concern is about our CNN example (Sect. 7). Its scope within the paper is to showcase the possibility of including deep architectures within our GLM framework. Deep CNNs have already been shown to have great performance on retinal recordings (Ref. 9), and it is not our scope to deepen these investigations, nor to claim that our bare CNN (excluding the GLM coupling filters) is better than those already presented in the literature. We will anyhow add a supplementary section with more explanation and details about our CNN model. For the same reason, we did not find it necessary to develop a whole CNN model for the moving bar that would have distracted the reader from our main messages. Minor points. As the reviewer supposes, stimulus filters do not generalise across stimuli, neither are they expected to do so. They are even of different dimensions: 3d for white noise, 2d for moving bar. We acknowledge that one can misunderstand this from the abstract, and we will modify it promptly. In addition, the analysis of Fig.3D is not done for the case of Fig.4 because when simulating the model with the couplings inferred via maximum log-ℓ (as opposed to our strategy) we observed self-excitation during all runs (Fig. 4E), and this prevented us from estimating predicted noise correlations. Lastly, for evaluating our prediction of noise correlations, the coefficient of determination (CoD) i is computed as 1- var(error) / var(data), and not as ρ^2. We will add these explanations.

Reviewer 3. Additional fine tuning. Good point. The correction of Eq. 4 plays the role of an additional fine tuning after the separate inference of stimulus and couplings filters. As explained above, Eq. 4 is indeed an approximation. An additional, not approximated fine tuning could consist in an inference of the stimulus filters after freezing the couplings network from step-2. We tried, but this approach came with no improvements and at a cost of more complex and time-consuming inferences. We therefore avoid including this additional step. We will add a comment in the discussion. Finally, we agree that including more analysis on deep architectures within the GLM framework opens for novel and interesting investigation. This is what we plan for future works, but it lies beyond the scope of the present work.

Reviewer 4. We believe the reviewer is missing a large part of our results, by focusing their criticisms on runaway instabilities. While important, these instabilities are not the main focus of our manuscript. In addition to runaway effects and the inclusion of CNNs into GLMs, we have focused on preventing stimulus biases in the GLM inference (Sect. 4, 5 and 6). These biases are particularly detrimental for complex stimuli: they prevent the correct estimation of noise correlations (Fig. 1G&E), and reduce the generalisability of inferred parameters (Fig. 4E) . Our two-step inference is constructed for dealing with these problems and is shown to solve them (Figs. 3 and 4). It accounts **also** for runaway effects, showing that the model inferred with our strategy is robust and powerful. Prior works. Gerhard et al. (2017) focuses only on single-neuron GLMs and lets for future developments the case of GLMs with a coupling network. That work surely merits a citation, but a quantitative comparison with our method for neuronal populations would not be meaningful. In addition, we never claim that our method provides better results than the other discussed approaches (Refs. 12 and 13), but simply that they are computationally harder than our method. Recurrent CNN. We agree that CNNs equipped with recurrent layers can account for temporal noise correlations between neurons. In fact, in a broad sense our method falls in this class of models. In the text we were referring to non-recurrent CNNs, yet we acknowledge the reviewer's criticism and we will happily add a comment and some citations to avoid additional misinterpretations. Repeated data. As clearly stated in the discussion, we acknowledge that the necessity of repeated data is a strong limitation of our method. For all the computational experiments we present we used only 2/3 mins of recordings, and this was somehow the lower-bound for our approach to work. However, let us stress that 2/3 mins does not represent a prohibitive cost within experiments that can last for two hours at least.