Dear editorial team, dear reviewers,

Please find below our responses to the remaining comments of reviewer 1. Please note, we have not yet made any changes to the bioRxiv copy of the manuscript.

Reviewer #1:

The manuscript is largely unchanged from the original submission and I still have major concerns with some aspects of the analyses and clarity of data presentation.

1. I respectfully disagree with the authors' remarks regarding neuropil correction. Neuropil contamination is still an important factor in GCaMP recordings. Although GCaMP in intracellular, two-photon imaging inevitably collects some fluorescence from outside of the cells of interest due to the axial extent of the point spread function. These neuropil signals do indeed reflect local population activity if GCaMP is expressed using local viral injection. Correcting for neuropil contamination is particularly important when comparing the activity of different neuronal populations, such as cells expressing different IEGs in this manuscript. While it is true that there is no "perfect" approach for neuropil correction, not correcting at all not the lesser of two evils. Indeed, in my opinion correlation analyses without neuropil correction are far more suspect as neuropil contamination artefactually correlates all neurons. Keemink et al (Scientific Reports, 8, Article number: 3493, 2018) compared various approaches for neuropil corrections using both simulated and ground truth data (notably acquired at even higher resolution than in the present manuscript). Surprisingly, even the simplest approach of using a fixed correction coefficient (probably the most widely used approach in the literature) improves the correlation between two-photon fluorescence and ground truth spike trains.

While I would not insist that the authors apply neuropil correction in the main figures of the paper, all the major analyses should be repeated using neuropil correction and included as supplemental figures.

We will try to illustrate the problem of neuropil correction. First, the neuropil correction method described in Keemink et al., called FISSA that was developed in the Rochefort lab, is indeed one of the more sophisticated methods, which is very well implemented and documented. Thus, we have used FISSA here - the problems are even worse with simpler subtraction-based correction methods, as the mean population activity and correlation is a direct function of the scaling parameter.

Let's assume we want to quantify some metric based on correlations, e.g. the average correlation between neurons in a given recording (the problem is the same, independent of what the exact metric is). We can compute this with and without neuropil correction (using FISSA). The problem is that FISSA (as well as all other neuropil correction algorithms) has a free parameter (it is actually multiple parameters, but for the sake of simplicity let's focus on the most important one in FISSA NMF variant) the experimenter can choose freely. This parameter is called Alpha. Now we can compute population correlation as a function of alpha and compare this to the population correlation we get from the uncorrected data. One would expect the neuropil correction to reduce the average population correlation, as neuropil should primarily act to increase apparent correlations between neurons.

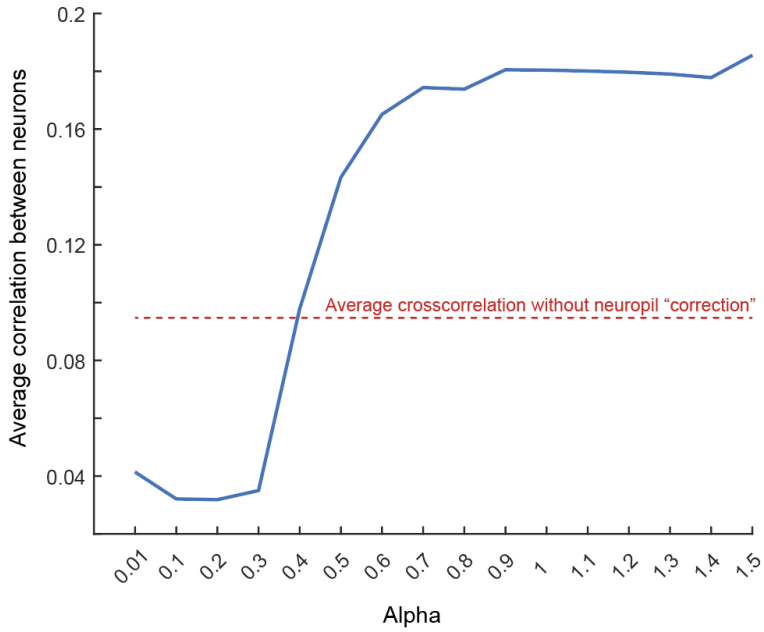

Figure R1. Average correlation between a population of neurons as a function of the alpha parameter of FISSA. The red horizontal line marks the correlation of the same neurons without neuropil correction.

What we find however, is that the average correlation between neurons varies by a factor of 5 (between 0.035 and 0.185) as a function of the correction parameter alpha (Figure R1). More worryingly, we can get the average correlation to decrease and increase away from the uncorrected correlation (0.095 here). Note, the range of alpha, that the original paper claims is stable, is: [0 0.5]. Well within this range we can choose the correction parameter such that we get the same result as without neuropil correction (or any other value between 0.03 and 0.16 for that matter).

Now, let's look at why this happens. In short, the algorithms are not as stable as advertised on uncalibrated data. We chose a random neuron from the dataset used above that would exhibit a pattern of activity one might suspect as neuropil contamination (Figure R2 – top trace). Plotting the different FISSA corrected traces clearly shows that A) the algorithm is not stable in the range of alpha between 0 and 0.5 on these data, and B) we can both dampen (0.01 or 0.1) or amplify the signal (0.5) as we choose. And worse, for certain values of alpha (0.03 and 0.04) we get traces that no longer look like fluorescence signals. In addition to this, the corrected dF/F values vary erratically, from 1.4 in the original data, to over 20 at alpha 0.3. This is only an example neuron, but the pattern is relatively consistent across the dataset, as is apparent also in the systematic change observed in Figure R1. And we will spare the reviewer the carnage that occurs when we now use a spike estimation algorithm (using CASCADE: https://www.nature.com/articles/s41593-021-00895-5) on these neuropil-corrected traces. On the example traces shown in Figure R2, we can get the total number of spikes detected to vary between 0 and over 2500.

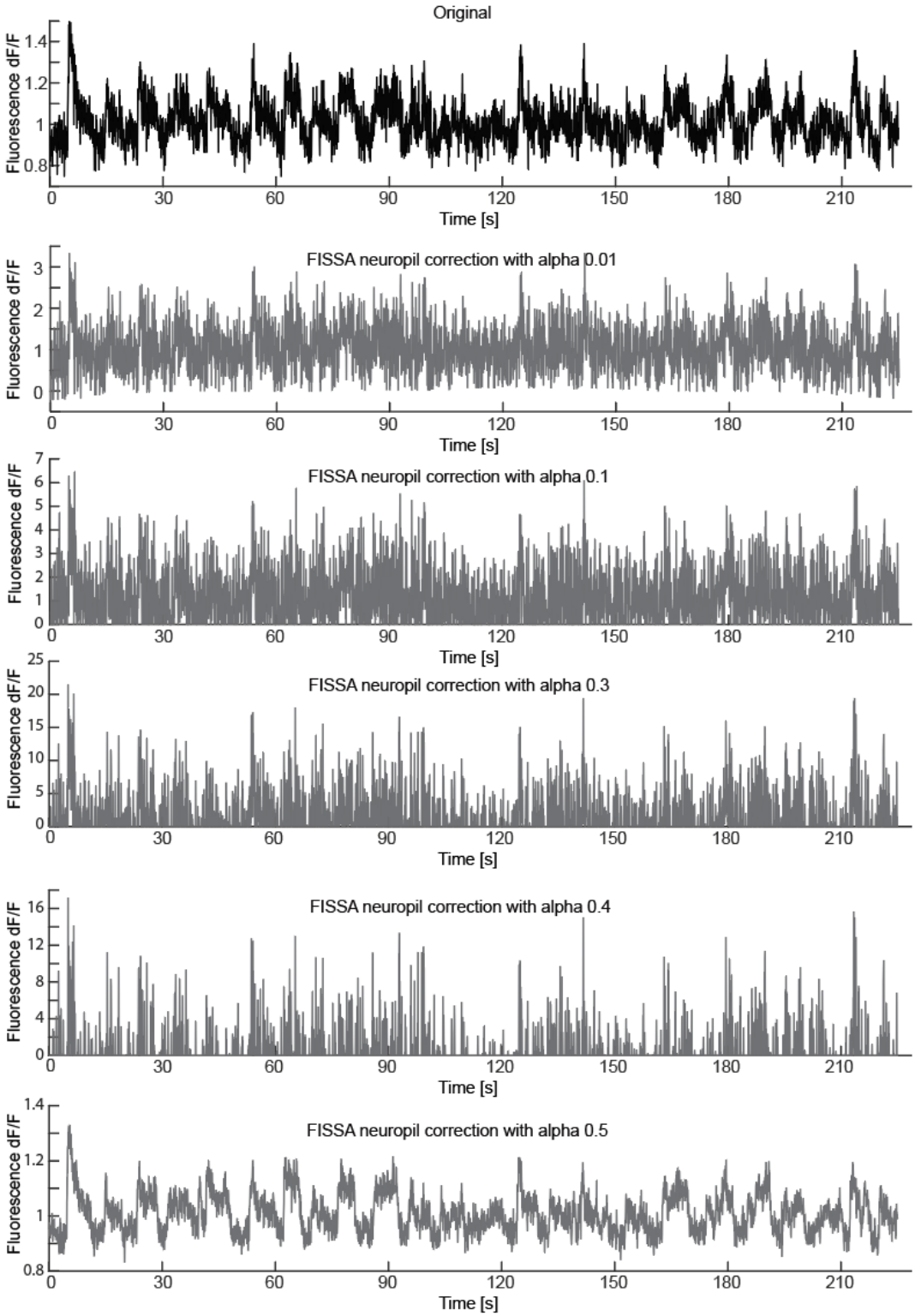

Figure R2. The dF/F traces of a sample neuron before neuropil correction and neuropil corrected with different choices of the alpha parameter using FISSA, as indicated in the panel titels.

What has become an unhelpful practice, is to include statements of the form "neuropil correction did not change our results" in manuscripts. Looking at Figure R1, it becomes apparent that for many analyses this statement must be true for a specific choice of alpha. Without specifying the details of what the authors did, this becomes meaningless. And worse, to a reader who has not worked with actual data and tested the effect of different parameter settings on their own data – this is uninterpretable.

For these reasons we will not neuropil correct the data, or include a subset of "corrected" analysis, or include any statement of the form "neuropil correction did not change our results". Not because it is difficult, or more work, but because we think it is bad practice to do so.

2. Figure 1E-J show positive correlations for both positive negative lags. I was referring to the negative lags, where IEG expression precedes activity measurement, as acausal. By baseline activity, I did not mean the raw F0 value but the average dF/F in the first dark recording preceding visual stimulation. If a subset of neurons has consistently high dF/F values and also shows above average levels of IEG expression, it will introduce a positive correlation between dF/F and IEG expression for all time lags. Correcting for baseline dF/F at the start of the experiment would help avoid this and reveal the relationship between IEG expression and changes in neural activity.

What is shown in Figures 1E-G is a cross correlation. We might be misunderstanding what the reviewer is proposing, but correlations are invariant to additive (or multiplicative) changes. Thus, no baseline correction constant to all time points will change anything in this figure. The width of cross correlations is typically directly related to the autocorrelation width of the two variables being cross correlated. In our case, the width of the cross-correlation function is determined by the comparably slow change of IEG expression levels. The reason the correlation remains positive even for "acausal" time lags, is that the IEG expression levels change with a time scale somewhere on the order of 3 to 6 hours.

3. The direct comparison between conditions 1 and 3 in Figure 2B is valid but the p-value reported for the negative slope is meaningless due to strong interdependence between time points.

We may be misunderstanding. Is the reviewer questioning the use of regression for linear trend analysis? If so, please explain. However, if the reviewer is unfamiliar with the idea of linear trend analysis, this is relatively standard practice, and the answer to the question of how a variable X changes with (typically) time (e.g. "Did global average temperatures increase over time" – on a side note, the answer to that is yes – stop eating meat). By design, the different time points are not independent measurements. We are testing the hypothesis of whether there is a linear trend in the mean activity levels after first exposure to light.

4. I had suggested that the authors should plot the distributions of individual points making up the averages in Figure 3. Although individual data points are shown in Figure 4A, the scatter plot is very difficult to parse as the data points heavily overlap. Please add some histograms to summarise the results either in Figure 3 or Figure 4.

As explained in the previous round, we prefer not to alter the way we show the data, unless there is a good reason for why that would be necessary. Please note, none of our arguments relate to shape of these distributions. Our claims are based on differences between the three populations which are visible best in the colored scatter plot (more blue vs. more green, etc). To plot histograms, we would need to add separate histograms for all three populations of neurons. And, as in Figure 3, the quantification is done exclusively on the means shown in the inset.