We thank the reviewers for their constructive comments on our submission. Below we address the raised concerns and include clarifications as suggested.

Experimental data. All reviewers noted a lack of experiments conducted on real data as a weakness of the work. We agree. As R4 suggests, we are ourselves working to validate these ideas experimentally. But there are two serious difficulties with using existing data in our context: **First,** because ground truth connectivity is almost never known for real data *in vivo*, most previous work, like ours, evaluates correctness on synthetic data, including (1; 2). **Second,** since we are modeling responses to stimulation of specific neurons, we require a dataset comprising similar manipulations. We know of no good benchmark data for these purposes, though we aim to generate them in future. Thus, our immediate follow-up plans involve performing simulations in more biologically plausible spiking networks, including nonlinear effects of photostimulation that more closely match biological responses (3; 4). This will allow us to directly compare our model to parametric approaches like (2; 5), since we agree with R1 that there are indeed likely to be cases and datasets in which these models outperform ours.

Relation to compressed sensing. We thank R3 for pointing out the highly relevant paper (1) and apologize for the omission. We now include this work (along with (2) in our discussion of inference from observation data using parametric models). While the manuscript already discusses relationships between our approach and one-bit compressed sensing, we had missed this work. **Key differences** between the present work and (1) include: **(a)** For speed and scalability, we focus on recovering binarized (present/absent) connections, not full weights. In practice, this may be all experiments require, and when it does not suffice, our approach may be used to rapidly pre-screen connections before performing more focused testing (an approach also suggested in (1)). **(b)** While the CoSaMP approach of (1) assumes a *known* level of sparsity, our method does not. In fact, our incorporation of uncertainty in weights allows for optional stopping. **(c)** We demonstrate scaling and speed necessary for implementation in the online setting. We plan to more fully discuss all these issues in our revised manuscript.

Theoretical note. In comparison with compressed sensing, we note an interesting connection: while our $\mathsf{a}_t = \|\mathsf{w} \odot \mathsf{x}_t\|_\infty$, the equivalent linear predictor for 1-bit CS is $\|\mathbf{w} \odot \mathsf{x}_t\|_1$, and our relaxed a_t is constrained to lie between these two. This raises the possibility that a generalization of our model might be able to interpolate between the two approaches.

Stimulation types. R3 noted as a weakness that we only consider randomized stimulation groups. This is incorrect. In our experiments, we also consider selecting subgroups adaptively by choosing to stimulate neurons with maximum marginal uncertainty (cf. Figure 3 in main text). As we show, this results in performance improvements over fully randomized stimulation.

Our contributions. While there has been much previous work on inferring network connectivity, as all reviewers note, the present work also contains novel methodological advances of broader interest, including: **(a)** the first application of group testing to network inference in neuroscience; **(b)** a novel relaxation of group testing, along with an equivalence to variational Bayesian inference; **(c)** a fast dual decomposition algorithm and GPU software implementation that makes online inference practical in large networks.

References

[1] Hu, T., A. Leonardo, and D. B. Chklovskii. "Reconstruction of sparse circuits using multi-neuronal excitation (RESCUME)". In Advances in Neural Information Processing Systems 22, pp. 790-798 (2009).

[2] Fletcher, A. K., S. Rangan, L. R. Varshney, and A. Bhargava. "Neural reconstruction with approximate message passing (NeuRAMP)". In Advances in Neural Information Processing Systems 24, pp. 2555-2563 (2011).

[3] Charles, A. S., A. Song, J. L. Gauthier, J. W. Pillow, and D. W. Tank. "Neural Anatomy and Optical Microscopy (NAOMi) Simulation for evaluating calcium imaging methods". bioRxiv:10.1101/726174 (2020).

[4] Luboeinski, J., and T. Tchumatchenko. "Nonlinear response characteristics of neural networks and single neurons undergoing optogenetic excitation". Network Neuroscience Advance publication. doi.org/10.1162/netn_a_00154 (2020).

[5] Aitchison, L., L. Russell, A. M. Packer, J. Yan, P. Castonguay, M. Hausser, and S. C. Turaga. "Model-based Bayesian inference of neural activity and connectivity from all-optical interrogation of a neural circuit". In Advances in Neural Information Processing Systems 30, pp. 3486-3495 (2017).