We thank the reviewers for their insightful comments and suggestions. We first reiterate the main goal and contributions. We asked (a) if having a $1/n$ neural code make neural networks more robust, and (b) how does the neural code employed by the intermediate layers affect the robustness. To answer these neuroscience-inspired questions on neural representation, we developed a pedagogical spectral regularizer that encourages an $1/n$ eigenspectrum on artificial networks. We demonstrated that networks with a $1/n$ eigenspectrum were more robust (sec 4.1). We provided empirical evidence that the neural representation employed by intermediate layers have a drastic affect on the robustness of the network regardless of the eigenspectrum of the last layer (sec 4.2, 4.3). We emphasize that the goal of this work is not to get SOTA performance on adversarial robustness or on other computer vision tasks, nor to design a practical training scheme. Rather, **our analyses elucidate the role of $1/n$ eigenspectrum observed in biological neural networks and also serve as inspiration for the design of future deep learning architectures; this is an exciting avenue of research that we leave for future work**.

As pointed out by R1,R2 & R3, our experiments were only run on MNIST. This was a deliberate choice as using MNIST has many advantages: 1) its simplicity makes it easy to design and train highly-expressive DNNs without relying on techniques like dropout or batch-norm, and 2) the models were able to be trained using a small learning rate, ensuring the efficacy of the training procedure detailed in section 3.2. This allowed us to isolate the effects of a $1/n$ neural representation, which may not have been possible if we use a dataset like CIFAR-10. We agree that the results would be bolstered if it were run on a natural image dataset like CIFAR-10 and leave that for future work but we note that the power-law behavior in rodents found by Stringer et al. was insensitive to the particular statistics of the input visual stimuli and was instead determined by the manifold dimension of the input. While the manifold dimension, d, of natural images is most likely larger than that of MNIST, both have manifold dimension much larger than 1; thus $\alpha = 1 + 2/d \approx 1$ for both MNIST and a dataset of natural images.

Per R1, we evaluated the robustness of networks on white-noise corrupted images where, in the interest of space, we only showcase the results of the networks from section 4.3 (Fig. 1). Having an $1/n$ eigenspectrum leads to an increase in robustness compared to their vanilla counterparts and for CNNs leads to **networks that are more robust than the Jacobian-regularized networks.** We would like to draw the attention of R5 to this particular case.

@R1: *"generic whitening is used for [. . .] deep networks, while batch normalization (BN) is only used for the shallow network . . . "* We apologize for the confusion. To clarify, the whitening employed in section 4.2 is used to investigate the importance of intermediate representations. Whitening leads to a flat eigenspectrum which is the worst case scenario under the theory of Stringer et al., thus we whitened only the second hidden layer in the networks to see how this would affect the robustness of the network. Only the networks that end with a `-Wh` in figure 4 are the ones whose intermediate neural representation was whitened. BN was only used for the shallow neural networks in section 4.1 as we found that it helped the spectrally-regularized networks reach $1/n$ faster than the networks without. We will make sure to clear this up in the updated manuscript.

@R2: *"Evaluating the networks on other vision tasks."* We agree that evaluating the networks on other computer vision tasks is interesting but it is outside the scope of the paper and we leave it for future work. @R2: *"...puzzled by the choice of regularization function . . . "* According to the theory developed by Stringer et al., having $\alpha < 1$ leads to undesirable properties. Thus, the goal of the regularizer is to get the eigenspectrum as close as possible to $1/n$ without going over, and the second term in the regularizer was added to heavily penalize eigenspectrum with an $\alpha < 1$.

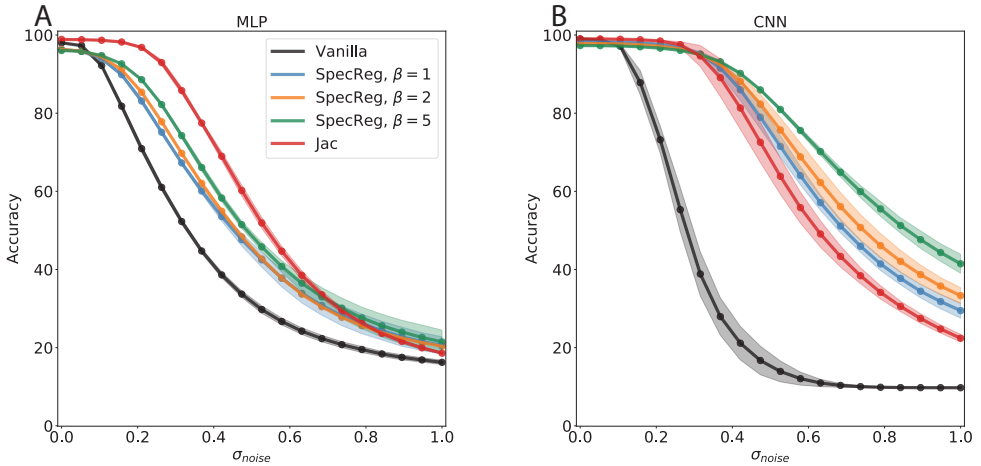

Figure 1: $1/n$ neural representation leads to robustness against **white-noise corrupted images**. The SpecReg networks correspond to the networks shown in section 4.3 where all the hidden layers were regularized. SpecReg is more robust than the Jacobian regularizer for the CNN.