We thank all three reviewers for their constructive and valuable feedback. They found our paper to be a "very interesting idea" (R1), to be a "good effort towards bridging the fields of neuroscience and machine learning" (R3), and to cover a "very meaningful topic" (R4). Their main concerns are to clarify experimental and numeric details and choices, as well as the implications for neuroscience and ML. We think that our comments below and several new analysis which we will include in the final version will clarify all open questions, and address most requested improvements.

R1: Principles behind the success of neurally regularized ML? We agree that this is *the* central question, and this is ongoing work. We do not have a final answer yet, but we will discuss some hypotheses in the final version.

R1: Training with the regularization alone. We performed related experiments in which we increased the relative importance of the similarity regularization over the classification loss (α in Eq. 7). Stronger regularization yields greater robustness, at the cost of worse classification on clean images. We will include these results in supplementary material.

R1: Consider testing semantically relevant perturbations. We fully agree. We did test models with/without neural regularization on a great variety of distortions (Fig. 1b). In all cases we find a better robustness of the regularized network, although to varying degrees depending on the distortion.

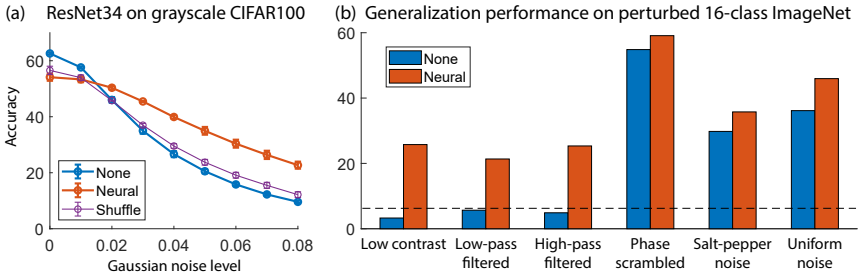

Figure 1: (a) Robustness test of ResNet34 models on CIFAR100. (b) Generalization result to naturalistic perturbations on 16-class ImageNet dataset.

R1: Why mice? While mice indeed do not have the sophisticated visual systems of primates, vision is still one of their major sensory inputs. Experimentally, mice allow for genetic tools for large scale recordings ($\sim$8000 units).

R3, R4: Selection of model layers to be regularized. Though V1 responses are thought to encode low-level features, there is no principled way to determine *a priori* which single model layer corresponds to V1. Thus we combine similarities from uniformly located layers (1, 5, 9, 13, 17 in ResNet18) flexibly without assigning to a specific one. Our model learns which are the optimal layers to regularize and find typically one lower layer.

R3: How can a predictive model trained on 100 images generalize to 5000? There is some misunderstanding here: the predictive model is trained on 5000 non-oracle images with the raw responses ρ_{ai} as targets. The 100 oracle images are used for evaluating the predictive model only (*e.g.* Fig. 1D in main submission).

R3: Correlation measure. The predictive quality weights v_a (line 96-97) are the correlation coefficients estimated over the 100 oracle (test) images. They bias the regularization toward similarities from better predicted neurons.

R3: Influence of prediction performance. The used models are highly predictive and state of the art ([13, 14], Ecker2019 ICLR), and we find their denoising effect to be crucial to reliably transfer robustness from experimental data.

R3: Other datasets and architectures? We reproduced our results on ResNet34 on CIFAR100 and a subset of ImageNet [3] (Fig. 1). We will include the results in the final version.

R3: You used 5000 images, not millions. We are referring to all pairs of images and the corresponding similarity targets, which are around $\frac{1}{2}5000^2 \approx 12.5\text{M}$ training samples.

R3: Use this technique without new experiments? Our predictive model generalizes across stimuli [13], letting us compute similarities without experiments. Future work on understanding why neurons improve robustness will also provide design principles without new experiments.

R4, R1: Previous work. To the best of our knowledge, we are the first to demonstrate that similarities derived from cellular level neural activities improves robustness in machine learning of any sort. Other work used similarity based regularization or fMRI activity, and we cite them in line 120 (Refs. [16, 17]). We will include a reference to Poggio.

R4: Why gray scale images? Mice are not sensitive to the colors relevant to human vision.

R4: Oracle images? Neural responses to the same stimulus are not exactly the same due to noisy responses. To estimate the reliability of each neuron we use randomly picked (oracle) images which we presented repeatedly during the experiment. When regularizing, each neuron contributed to the similarity measures according to its reliability.

R4: How many neurons? We are currently investigating this. However, we find the similarity matrices to be robust w.r.t. to the selection of neurons (Fig. 1 in main submission).