Inter-animal transforms as a guide to model-brain comparison

Abstract: To address the question of how to compare DNN model activations to brain data, we investigate what transforms best describe similarity in neural activity *in the same brain area between conspecifics*. We expect neural responses to be functionally highly similar within a species (since we expect findings to generalize across animals). What kind of transform will make such similarity most evident? That is, under what kind of transform are conspecifics' neural responses highly similar to each other? Researchers often default to linear regression as a reasonable transform class for measuring neural response similarity. We propose an improved transform class that uses a generalized linear model (GLM) whose noise matches the approximately Poisson noise in the neural data, and whose non-linear link function is akin to the activation function of a biological neuron. Incorporating these biologically motivated constraints into the inter-animal transform class substantially improves similarity scores compared to linear regression. We then build a DNN model of mouse visual cortex that swaps out ReLU activations for a more biologically plausible softplus activation function, combined with Poisson noise, to produce activations that are more similar to neural responses. We find that a Poisson GLM whose link function exactly matches the model activation function again yields the highest similarity scores between different randomly seeded instances of our softplus models. This result gives mechanistic insight into why the best performing animal transform class has a non-linear link as well as Poisson noise structure. Moreover, we show that our Poisson GLM not only achieves higher similarity scores for the same layer between model instances, but also scores activations in model layers that are physically far apart as highly dissimilar to each other. Finally, we estimate the number of neurons and number of stimuli that would need to be recorded to accurately estimate inter-animal similarity.

Significance: How to assess similarity in neural responses between a DNN model and the brain is an important methodological question for computational neuroscience. For example, deep convolutional neural networks are now widely touted as the most accurate models so far of a number of neural systems, such as the primate ventral visual stream and the mouse visual system. This conclusion is based on prior assumptions about how to measure model accuracy, which in many cases simply involve linearly regressing neural responses on model features. Is linear regression the most appropriate way to assess model-brain similarity, and more generally, what would a good measure of model-brain similarity be?

To answer this question, we investigate *empirically* what type of mapping (linear or otherwise) is most apt for predicting one animal's neural responses based on a conspecific's neural responses, and use *that* as our measure of model-brain similarity (cf. Cao and Yamins, 2021). The inter-animal similarity score according to this empirically validated transform is a kind of noise ceiling, i.e. the maximum level of similarity that any computational model of that species can hope to attain, according to this transform.

Our work is of interest not only to computational neuroscientists, but also to systems neuroscientists, in three ways. First, we provide evidence that the best transform class we've found so far depends on the neuronal activation function, as well as the structure of the noise in the neuron's spiking activity. As a result, more precisely identifying the best transform class between animals could provide information about the activation functions and noise characteristics of the neurons, and conversely, incorporating information from systems neuroscience about the neural activation function can lead to a better estimate of the transform class. Second, accurately modeling similarity and dissimilarity in neural responses between conspecifics is important for systems neuroscience, and our nonlinear transform based on a Poisson GLM provides some guidance on how to measure inter-animal variability. Third, we provide estimates of how much data (stimuli, neurons) would need to be collected to reliably estimate the inter-animal variability according to a linear transform (and in the future will provide these estimates for our non-linear transform, which is more sample efficient), thus providing guidance for future experimental work.

In overview, we show that on the mouse visual cortex Neuropixels recordings from the Allen Brain Observatory Visual Coding Dataset (de Vries *et al.*, 2020; Siegle *et al.*, 2021), a Poisson GLM is the best transform class between conspecifics (Fig. 1A). We design a more biologically accurate model of visual cortex (Fig. 1B) and show that the same Poisson GLM achieves the highest same-layer similarity scores (Fig. 1C). For different transform classes, we measure both same-layer and cross-layer similarity between models trained with different weight initializations and data orderings and emphasize a second desirable property of a good transform class: that it also preserves separability between distant layers (i.e. minimizes cross-layer similarity) (Fig. 1D). Lastly we estimate the amount of data needed to accurately estimate inter-animal variability (Fig. 1E). Taken together, our results encourage looking beyond linear mappings when comparing models to neuronal data and in particular to consider transform classes which resemble the neuronal mechanism underlying the data.

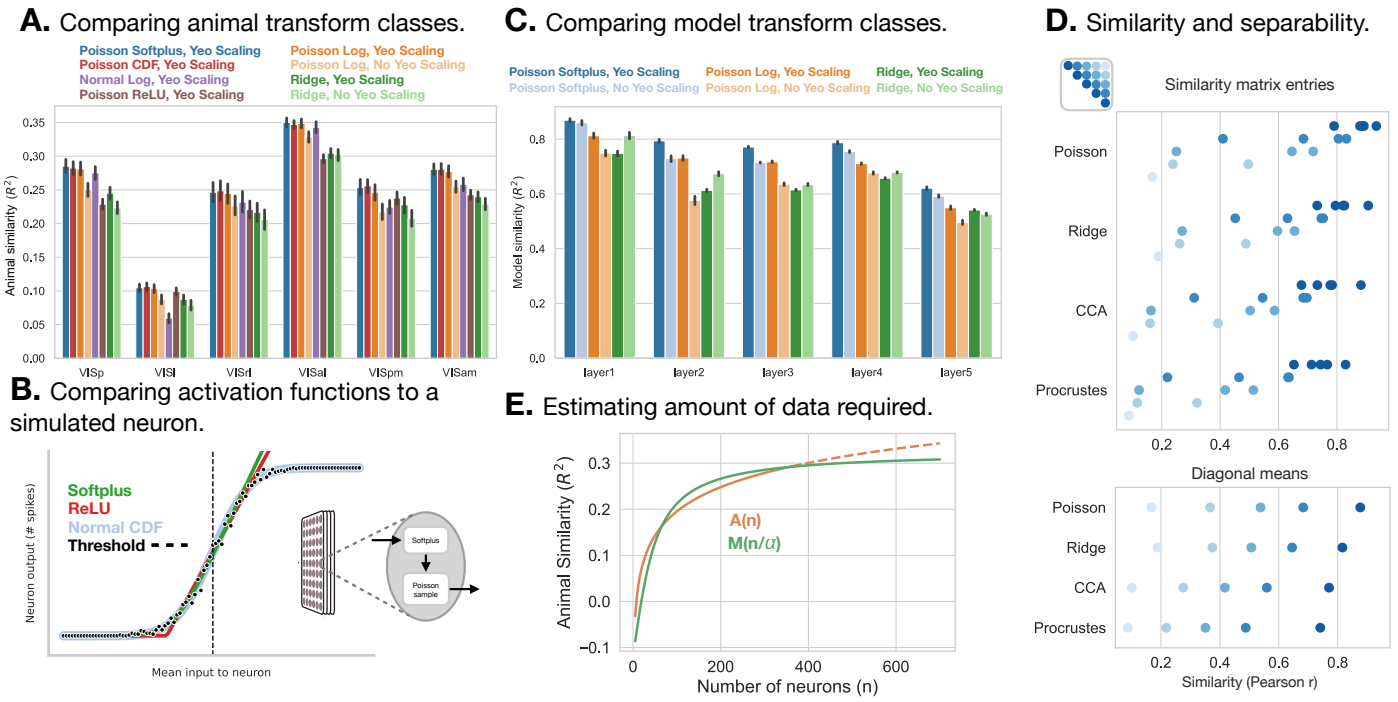

Figure 1: A Poisson GLM with a Softplus link function shows a substantial improvement in transform similarity compared to linear mappings. A. A Poisson GLM using a link function based on the Softplus activation function (dark blue bar, far left), combined with Yeo-Johnson scaling on the source features, substantially outperforms Ridge regression in terms of inter-animal similarity scores. Three components contribute to the higher scores. First, since neuronal spiking activity can be approximately modeled with a Poisson distribution, incorporating this noise structure into the transform between animals improves its predictive accuracy, especially for small datasets. Second, Yeo-Johnson scaling transforms the source features so that they are approximately normally distributed. Neural responses tend to have skewed distributions, and skewed predictors are often less effective when using linear regression or GLMs. Third, using a non-linear link function such as the log link, softplus-based link, or Normal Cumulative Distribution Function (CDF), leads to better results. **B.** We simulate a neuron's output (spike counts) as the result of repeatedly sampling a stochastic process in which the neuron spikes if its input exceeds the neuron's threshold (dotted line). The total input to the neuron is the aggregation of noisy inputs from thousands of other neurons, so we model it as a Gaussian variable. A noisy input can sometimes exceed the neuron's threshold and cause it to fire, even when the *mean* input is below threshold (the unsaturated, fluctuation-driven regime, left of the dotted line). This simple model yields a sigmoid activation function (the Normal CDF). However, since the cortical neurons we measure are mostly unsaturated, we can approximate their activation function with Softplus or ReLU, with Softplus providing a better approximation. This motivates using the Softplus function not only as the link function in our GLM, but also as the activation function in our models. We modify an Alexnet model of mouse visual cortex introduced in Nayebi *et al.*, 2022 by replacing ReLU with Softplus in each layer, and using the Softplus activation as the mean of a Poisson distribution from which we sample the noisy activations. The model activations can thus be interpreted as the number of occurrences (or spikes) in a given time period. **C.** A GLM whose functional form matches the mechanism that generates the model's activations achieves the highest transform similarity. The best performing transform class for the models matches the best performing transform class for the animal data (see blue bar in Fig. 1A). **D.** Same-layer and cross-layer similarity scores can typically be presented in a symmetric matrix and visualized as a heatmap (top left). In the upper panel, we instead display the scores in a strip plot to highlight the differences in numerical values of the entries in the upper triangular portion of the matrix. The darkest entries represent same-layer similarities (i.e. entries on the matrix diagonal). Lighter colors represent cross-layer similarities, with lighter shades corresponding to entries further away from the diagonal of the similarity matrix (layers that are further apart from each other). The bottom panel shows the average value for all entries in the k-th diagonal of the similarity matrix, for each value of k. Intuitively, a good transform class should score high on same-layer similarity, i.e. the darkest dot should be as far right as possible. In assessing layer separability, we want the lighter dots to be as far left as possible, while still being spaced apart. **E.** For any given transform class, we can quantify the effect on similarity of subsampling the number of stimuli or units the transform class is trained on. We can fit a curve to extrapolate performance as a function of number of model units and calibrate it against the animal curve to estimate the number of neurons we'd need to record from to maximize the transform class's performance. For linear mappings, our estimates indicate we would require data for at least 500 stimuli and 500 reliable neurons for such a goal. We expect these estimates to be lower for the best transform class, and we'll report these numbers soon.