# Predictivity and specificity for model-brain alignment methods

**Imran Thobani (ithobani@stanford.edu), Javier Sagastuy-Brena (jvrsgsty@stanford.edu),**
**Aran Nayebi (anayebi@mit.edu), Rosa Cao (rosacao@stanford.edu), Dan Yamins (yamins@stanford.edu)**
Stanford Neuroscience and Artificial Intelligence Laboratory, Stanford University, CA, USA

## Abstract

**The appropriate methods for aligning neural network models to the brain remain controversial. Ideally, a good alignment method should be powerful enough to enable accurate predictions of neural responses under a mapping from model units to neurons, while also being specific enough to distinguish the target system (e.g. a particular brain area) from other systems. It has generally been assumed that the goals of predictivity and specificity are in tension with each other, with methods that severely restrict the possible relationships between model and target being better for specificity, and more flexible methods yielding higher predictivity. We show that this apparent tension does not in fact exist. Fundamentally, this is because specificity requires not only distinguishing response patterns from different brain areas (i.e. separation), but also recognizing response patterns from the *same* brain area as being similar across subjects (i.e. identification). Taking this into account, we find that relatively flexible methods, like linear regression, can exhibit greater specificity compared to stricter methods, while also enabling better predictions. Motivated by the idea that the correct balance between strict and loose is manifested by the empirical relationships between subjects in a population, we introduce an alignment method that incorporates known aspects of the biological circuit, further improving predictivity without reducing specificity.**

**Keywords:** alignment; mappings; neural predictivity; neural networks; mechanisms

## Introduction

Aligning neural networks to the brain has been challenging because it has been unclear what the criteria for good alignment methods are. Ideally, a good alignment method should succeed on two fronts. First, it should enable *accurate predictions* of neural activity, implemented via a mapping from model components to neural components that aligns simulated and real activity. Second, an alignment method should exhibit *specificity*, identifying response patterns from the same part (e.g. brain area or model layer) as being similar across different instances of the population, while distinguishing response patterns from different parts as being dissimilar.

It has been widely presumed that the goals of predictivity and specificity are in tension with each other (Ivanova et al., 2021). Intuitively, more flexible transform classes appear better for prediction, while stricter transform classes appear better for separation. To the extent that this trade-off exists, there is an inherent divergence between the goals of accurate prediction (e.g. building brain-machine interfaces) and scientific understanding (e.g. systems identification). Indeed, this assumption has had substantial influence on metric design, with researchers pursuing scientific understanding favoring stricter methods (Williams, Kunz, Kornblith, & Linderman, 2021) and those pursuing engineering applications favoring more flexible methods (Schrimpf et al., 2018).

However, the literature overlooks a crucial aspect of specificity: recognizing instances of the same type (e.g. of the same brain area or model layer) as similar. Indeed, an alignment method that indiscriminately separated *all* response patterns would be incapable of recognizing target systems of the same type (e.g. the same brain area and species) as similar to each other, and therefore would lack specificity. Considering both aspects of specificity suggests that rather than a trade-off, there is an optimal balance between strictness and flexibility, where we want the *narrowest* class of transforms that accurately maps responses between subjects for the same brain area. To better approximate this ideal, we propose a transform class that accounts for known aspects of the biological circuit, increasing predictivity without reducing specificity.

## Aligning subjects in a model population

We evaluate several alignment methods, including soft matching (Khosla & Williams, 2023), ridge regression, and Representational Similarity Analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008). Soft matching is a strict method that matches individual units with "soft" permutations. Because we formulate soft matching as a predictive model, we *can* evaluate it for predictivity. On the other hand, RSA does not provide a mapping, so cannot be evaluated for predictivity, only for specificity. We also introduce and evaluate methods that incorporate aspects of a biological activation function.

We first evaluate these methods on a simulated mouse population because, unlike neural data, we can sample responses for all units over arbitrarily many stimuli, allowing for clearer results. We use a modified AlexNet that predicts mouse brain responses under a linear mapping better than other models (Nayebi et al., 2021). The models have a smooth activation function (softplus) and Poisson-like noise to better match cortical response properties (Stevens & Zador, 1995; Dapello et al., 2020). To simulate different subjects, we vary the random seed controlling the weight initialization and training data order. Although our model population likely does not fully capture true inter-animal variability, we also evaluate methods on actual mouse data and obtain similar results.

We assess same-layer predictivity with the $R^2$ score on held-out responses, after fitting parameters to map one model instance to another. To evaluate *both* aspects of specificity, we compute the silhouette score (Rousseeuw, 1987), which is close to 1 just in case responses for different layers are separated much more than responses for the same layer. The silhouette score for response profile $i$ is:

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}$$

where $a(i)$ is the mean dissimilarity between $i$ and other response profiles for the same model layer, and $b(i)$ is the mean dissimilarity between $i$ and responses from the next most similar model layer. We compute the mean silhouette score over all model subjects and layers.

Ridge regression (green bars) achieves higher predictivity than soft matching (orange bars) (Fig. 1A). However, a

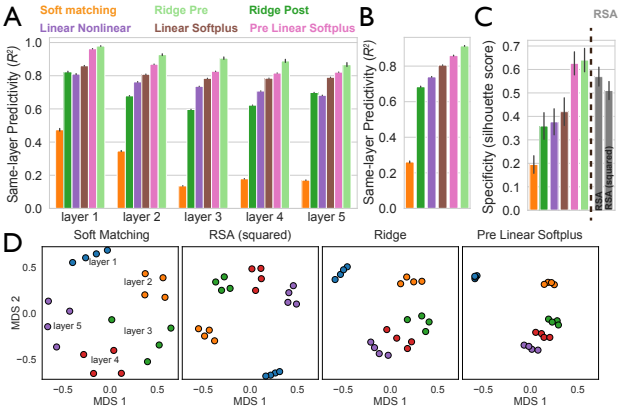


Figure 1: Model population. (A) Same-layer predictivity between model subjects. (B) Overall predictivity. (C) Silhouette scores. RSA scores are on a different scale from $R^2$, so we also consider squared RSA scores. (D) Multidimensional scaling of dissimilarity (1-similarity) between response profiles.

further substantial increase in predictivity occurs for ridge regression on *pre*-softplus activations (light green bars) at every layer. Thus, model responses converge after applying the filter weights, diverge after the activation function, and converge again in the next layer.

Rather than finding a trade-off, we find that increased predictivity can *improve* specificity (Fig. 1B, C) as long as inter-layer separation is maintained. For example, ridge (post-softplus) exhibits *more* specificity than soft matching (Fig. 1B). This is because ridge clusters same-layer responses more tightly than soft matching while maintaining inter-layer separation (Fig. 1D). Ridge (pre-softplus) achieves an even higher silhouette score, in line with its increased predictivity. In fact, a maximally specific transform class should achieve maximum predictivity for same-layer responses, while being as narrowly defined as possible.

To develop alignment methods that perform better on post-non-linearity responses (e.g. firing rates), we introduce transform classes that account for the activation function. Linear Nonlinear approximately inverts the non-linearity using Yeo-Johnson scaling and then uses a generalized linear model to apply a fitted linear mapping followed by a smooth non-linearity (the exponential). Linear Softplus precisely matches the softplus activation function, and Pre Linear Softplus maps *pre*-non-linearity responses of the source model to post-non-linearity responses of the target model (i.e. exactly rather than approximately inverting the softplus non-linearity). These methods improve both predictivity *and* specificity, with Pre Linear Softplus performing best (Fig. 1B, C).

## Aligning mouse subjects and models to mouse

We investigate how well our results generalize to a mouse dataset containing Neuropixels recordings for 31 subjects in response to 118 naturalistic stimuli, averaged over 50 trials (de Vries et al., 2020). With only about 50 neurons measured per subject and brain area, we pool N-1 subjects' neurons to evaluate same-area predictivity for a target subject. Overall, the rank ordering of alignment methods in terms of same-area predictivity is similar for the real population as for the simulated population (Fig. 2A). This helps validate our simulated population as a model of inter-animal variability, at least to some degree.

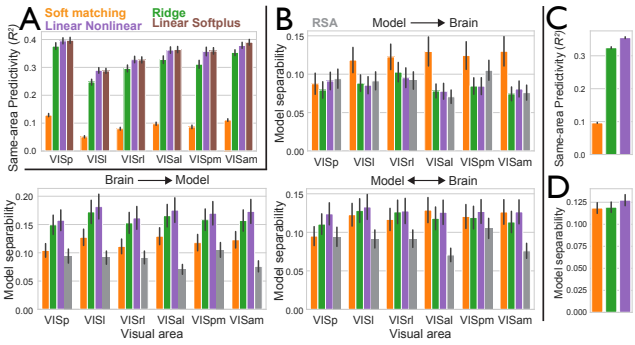


Figure 2: Mouse population. (A) Same-area predictivity between subjects. (B) Model separability with respect to brain similarity. (C) Overall predictivity. (D) Overall model separability (both directions).

When pooling across subjects, we cannot compute silhouette scores. We therefore assess specificity *indirectly* by considering the average difference between 4 candidate models in terms of assessed similarity to each brain area. A method with low specificity would not be able to differentiate models that are *more* similar to the brain from those that are *less* similar, and therefore would have low model separability.

We map model-to-brain as well as brain-to-model. For model-to-brain, soft matching separates models better, consistent with (Khosla & Williams, 2023), but for brain-to-model, Ridge and Linear Nonlinear separate models better (Fig. 2B). A possible reason is that model responses can have patterns not present in the brain data, and flexible mappings like Ridge or Linear Nonlinear may only detect such a discrepancy when mapping from brain to model. When mapping in both directions, model separability is as good for Ridge and ILNP as it is for soft matching. Overall, there is not a trade-off between predictivity and model separability (Fig. 2C, D).

## Conclusion

There is not a systematic trade-off between predictivity and specificity. In fact, both goals should be achieved by the *narrowest* class of transforms under which subjects' responses predict each other with high accuracy for that area. To better approximate that class, we introduce a method that accounts for the activation function, improving predictivity while maintaining specificity. Future research should investigate whether we can further constrain ILNP in a way that improves specificity without reducing predictivity.

## References


Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, *33*, 13073–13087.

de Vries, S. E., Lecoq, J. A., Buice, M. A., Groblewski, P. A., Ocker, G. K., Oliver, M., ... others (2020). A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature neuroscience*, *23*(1), 138–151.

Ivanova, A. A., Schrimpf, M., Anzellotti, S., Zaslavsky, N., Fedorenko, E., & Isik, L. (2021). Beyond linear regression: mapping models in cognitive neuroscience should align with research goals. *bioRxiv*. doi: 10.1101/2021.04.02.438248

Khosla, M., & Williams, A. H. (2023). Soft matching distance: A metric on neural representations that captures single-neuron tuning. *arXiv preprint arXiv:2311.09466*.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

Nayebi, A., Kong, N. C., Zhuang, C., Gardner, J. L., Norcia, A. M., & Yamins, D. L. (2021). Unsupervised models of mouse visual cortex. *bioRxiv*.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... others (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Stevens, C., & Zador, A. (1995). When is an integrate-and-fire neuron like a poisson neuron? *Advances in neural information processing systems*, *8*.

Williams, A. H., Kunz, E., Kornblith, S., & Linderman, S. (2021). Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, *34*, 4738–4750.