Model-brain comparison using inter-animal transforms

Imran Thobani (ithobani@stanford.edu)
Department of Philosophy, Stanford University, 450 Jane Stanford Way, Stanford, CA 94305
Javier Sagastuy-Brena (jvrsgsty@gmail.com)
Opal Camera, 143 2nd St, San Francisco, CA 94105
Aran Nayebi (anayebi@cmu.edu)
Machine Learning Department, Carnegie Mellon University, 5000 Forbes Avenue, 8th Floor, Pittsburgh, PA 15213
Jacob Prince (jacob.samuel.prince@gmail.com)
Department of Psychology, 33 Kirkland St, Harvard University, Cambridge, MA 02138
Rosa Cao (rosacao@stanford.edu)
Department of Philosophy, Stanford University, 450 Jane Stanford Way, Stanford, CA 94305
Daniel Yamins (yamins@stanford.edu)
Department of Psychology, Stanford University, 450 Jane Stanford Way, Stanford, CA 94305

Abstract

Artificial neural network models have emerged as promising mechanistic models of brain function, but there is little consensus on the correct method for comparing activation patterns in these models to brain responses. Drawing on recent work on mechanistic models in philosophy of neuroscience, we propose that a good comparison method should mimic the Inter-Animal Transform Class (IATC) - the strictest set of functions needed to accurately map neural responses between subjects in a population for the same brain area. Using the IATC, we can map bidirectionally between model responses and brain data, assessing how well the model can masquerade as a typical subject using the same kinds of transforms needed to map across animal subjects. We attempt to empirically identify the IATC in three settings: a simulated population of neural network models, a population of mouse subjects, and a population of human subjects. In each setting, we find that the empirically identified IATC enables accurate neural predictions while also achieving high specificity (i.e. distinguishing response patterns from different areas while strongly aligning same-area responses between subjects). In some settings, we find evidence that the IATC is shaped by specific aspects of the neural mechanism, such as the non-linear activation function. Using IATC-guided transforms, we obtain new evidence, convergent with previous findings, in favor of topographical deep neural networks (TDANNs) as models of the visual system.

Keywords: similarity scores, model-brain comparison, neural prediction

Introduction

Artificial neural network (ANN) models have been found to exhibit internal activation patterns that predict aspects of neural activity in a wide variety of brain areas (Zipser & Andersen, 1988; Olshausen & Field, 1996; Yamins et al., 2014; Storrs et al., 2021; Kell et al., 2018; Zhuang et al., 2021; Khaligh-Razavi & Kriegeskorte, 2014; Sussillo et al., 2015; Wang et al., 2021; Schrimpf et al., 2021; Mineault et al., 2021; Nayebi et al., 2021, 2024). These results naturally raise the question of whether these models can serve as mechanistic models of brain function, at least at some level of abstraction (Cao & Yamins, 2021; Kriegeskorte & Diedrichsen, 2016). Although a variety of methods have been proposed for quantitatively assessing neural response similarity between models and brains (Sucholutsky et al., 2023), there is little consensus on what the correct method is.

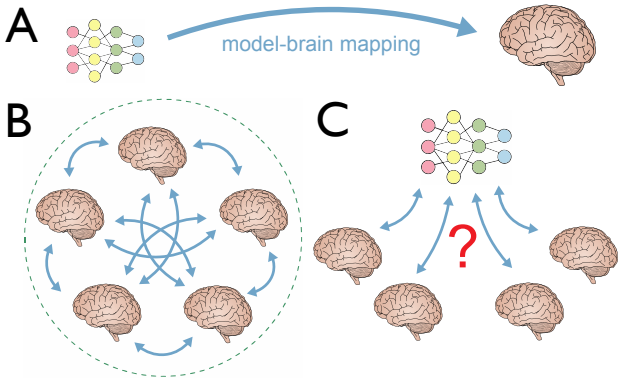

Figure 1: **Model-brain comparison using inter-animal transforms.** (A) Model-brain mappings: We seek a principled method for comparing model responses to brain responses using predictive mappings. (B) The Inter-Animal Transform Class (IATC) is the strictest set of functions required to map responses accurately between subjects in a population for a given brain area. (C) We propose to use the empirically identified IATC to map bidirectionally between a candidate model's responses and brain responses in order to assess whether the model can masquerade as a typical animal subject.

A particularly stringent approach to comparing models to brains would be to use 1-1 matching, i.e. attempting to bijectively identify each ANN model unit with a unique neuron in a target animal's brain. However, a key challenge complicating model-brain mapping is the fact that the target of modeling is not a single idealized brain, but rather a *population* of brains that are all somewhat different from each other. In fact,

inter-subject variability can be substantial – in humans, the estimated number of neo-cortical neurons can vary between subjects by up to a factor of 2 (Haug, 1987), and even the exact number of functionally identified brain areas can vary (Gao et al., 2022). As a result, 1-1 matching is likely to be problematic. A more sophisticated approach to model-brain mapping is therefore needed.

A first generation of approaches to this problem used linear mappings to compare models to brains (Yamins et al., 2014), but concerns have arisen that the linear mapping class is too flexible to strongly separate models of the brain (Kornblith et al., 2019; Ding et al., 2021; Conwell et al., 2022). More recent methods have thus focused on "stricter" mapping classes, such as soft matching (Khosla & Williams, 2023), that tighten the criteria for model-brain similarity while still allowing comparisons between different-sized populations of neurons (unlike 1-1 matching).

Here we develop the idea of the ***Inter-Animal Transform Class*** (IATC), a concept that has been introduced in philosophy of neuroscience to handle the problem of between-subject variability when building mechanistic models (Cao & Yamins, 2021). As defined in that work, the IATC is the strictest (smallest) class of functions that maps responses between any two subjects in a natural population, matching the same brain area across subjects with as high accuracy as possible (Fig. 1B).[1] Both the strictness and mapping accuracy criteria are important: a good IATC candidate must (by virtue of mapping accuracy) succeed in *aligning* same-area responses across subjects, while (by virtue of strictness) *separating* responses from different areas. This pair of desiderata naturally suggests a meta-metric of *specificity*, which evaluates the extent to which a mapping simultaneously achieves high cross-subject within-area identifiability while maintaining between-area separability (Fig. 2A). As an extension of specificity, we also consider whether similarity scores under a mapping method correlate with inter-area distances in a known hierarchy (Fig. 2B).

Though defined relative to a population of real individuals, the IATC can be used to compare artificial networks to brains. Specifically, we propose to use the empirically-identified IATC itself to map between models and brains – in effect, measuring how well the model can masquerade as a member of the population (Fig. 1C). A key implication of this IATC-based approach is that model-brain mappings should be performed *bidirectionally* between models and brain data, just as when comparing two brains to each other, rather than only mapping in one direction (from model to brain).

A primary challenge in applying the IATC is how to practically identify it for a given population. Ideally, we would estimate the IATC directly using large-scale optimization techniques applied to a massive neural dataset with many subjects. In the absence of the required data and techniques for doing so, here we instead evaluate a spectrum of well-known

[1] The maximum possible accuracy that is obtainable may be less than perfect because different subjects' neural representations can have different metamers (Feather et al., 2023), and therefore different neural encoding functions with different null spaces.

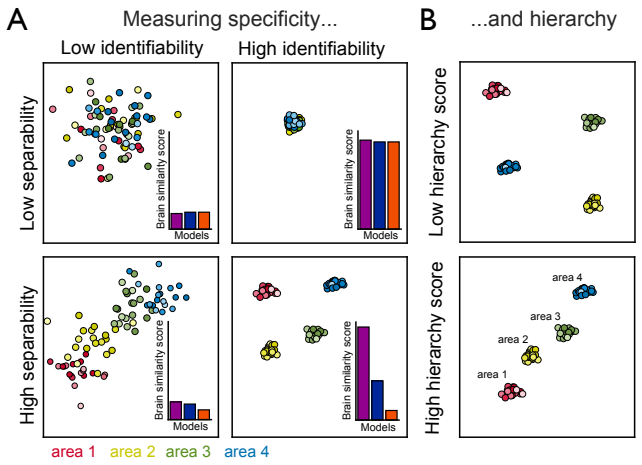

Figure 2: **Evaluating candidate transform classes for specificity and hierarchy.** (A) It is desirable for a mapping to achieve high *specificity* – simultaneously achieving within-area *identifiability* and between-area *separability* when mapping responses between animals. Each dot is a response profile for a particular subject and brain area. The schematic barplots represent likely outcomes for model separation when comparing models to brains. (B) To capture more graded relationships beyond specificity, we also look at the correlation between dissimilarity scores and distances in a known hierarchy.

methods, such as linear regression and soft matching, against the two basic IATC criteria: they should map responses across subjects within an area as accurately as possible, while separating responses from different areas.

We first evaluate candidate transform classes on a simulated population of artificial neural network models. Because we have complete knowledge of the network structure and can "measure" responses for all units over many stimuli, we are able to observe how the specific form of the activation function in the network shapes the relationships between model subjects' responses. This motivates a new transform class that maps responses across model subjects with close-to-maximum predictivity and high specificity, yielding a reasonable estimate of the IATC. We then evaluate these different methods on real neural datasets, including both mouse electrophysiology and human fMRI recordings.

Results

Testing candidate IATCs for a simulated population

We first evaluate transform classes on a simulated population of neural networks (Figure 3A) against the IATC criteria by testing for same-area predictivity as well as for specificity. Our simulated population consists of neural networks based on a state-of-the-art model of mouse visual cortex: an AlexNet trained with contrastive learning on 64x64 inputs (Nayebi et al., 2022). We further modified the model to use

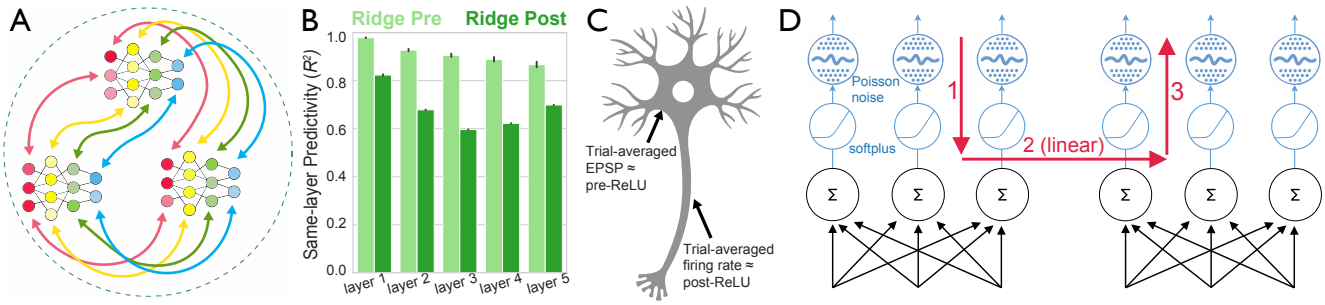

Figure 3: **Assessing same-area similarity in the model population reveals a "zippering" effect caused by the model activation function.** (A) We attempt to identify the IATC for a model population by first assessing within-area similarity when mapping between differently seeded model subjects. (B) Zippering effect: Ridge regression accurately maps pre-non-linearity responses, but not post-non-linearity responses, between subjects at each layer. (C) Post-non-linearity responses can be thought of as corresponding to firing rates, while pre-non-linearity responses can be thought of as corresponding to EPSPs (excitatory post synaptic potentials). (D) Inverse Linear Softplus: A schematic of a transform class that considers the effect of the non-linearity. Step 1 inverts the non-linearity to recover the pre-non-linearity activations of one subject, step 2 applies a fitted linear transform to predict the pre-non-linearity activations of the other subject, and step 3 re-applies the non-linearity to predict post-non-linearity activations.

a softplus activation function followed by Poisson-like noise to better mimic neuronal response characteristics. To generate a population of model subjects, we vary the random seed controlling the weight initialization and training data order. We map responses between subjects using 10000 activation patterns driven by ImageNet-validation stimuli (80/20 train-test split), evaluating the test R^2, median across target neurons for a given model layer, averaged in both directions and across all pairs of subjects.

The activation function has a substantial effect on same-layer response similarity between model subjects. Because a viable IATC candidate must map responses accurately across different subjects for the same layer, we first evaluate ridge regression, which has been widely used for neural response prediction (Canatar et al., 2024). Surprisingly, ridge regression achieves only moderate same-layer predictivity for intermediate model layers when mapping post-softplus activations between subjects (Fig. 3B). This raises the possibility that the IATC might require highly non-linear transforms such as those implemented by an MLP, which if true would suggest that model subjects trained from different random seeds are highly dissimilar in their learned representations.

However, we observe a "zippering" effect: at each layer, pre-non-linearity responses are close to linearly related between subjects, but the activation function disrupts these linear relationships for post-non-linearity responses, before the next layer's pre-non-linearity responses become linearly related again (Figure 3B). This effect suggests that the model subjects are actually similar, despite the apparent divergence suggested by the failure of ridge regression to map post-non-linearity responses accurately. Pre-non-linearity activations can be thought of as corresponding to trial-averaged EPSPs in real neurons, while post-non-linearity activations can be thought of as corresponding to trial-averaged firing rates (Figure 3C). Because EPSPs are hard to measure, we develop an IATC candidate that works for post-non-linearity responses.

Improving cross-subject mapping by considering the non-linear activation function. The results for ridge regression pre- and post-non-linearity suggest that an ideal IATC candidate must consider the effect of the non-linearity on the relationships between different subjects' post-non-linearity responses. We develop a transform class called Inverse Linear Softplus, which inverts the softplus activation function, applies a fitted linear mapping between the two subjects, and re-applies the softplus activation to predict the target subject's post-softplus responses (Fig. 3D). Building on an established framework of generalized linear models (GLMs) (McCullagh & Nelder, 2019; Chichilnisky, 2001), we use a GLM whose inverse link function is precisely matched to the softplus activation in order to fit the linear mapping and re-apply the softplus activation function. This yields substantially higher predictivity than ridge regression when mapping post-softplus activations (Fig. 4A).

In the case of real neural data, we may not know the exact form of the activation function, and we therefore develop versions of our transform class that attempt to approximately account for the activation function. Linear Softplus (unlike Inverse Linear Softplus) approximately inverts the activation function (step 1 of Fig. 3D) for the source model subject by using Yeo-Johnson scaling. Yeo-Johnson scaling applies a power transformation to make the post-non-linearity features normally distributed, and thus more closely resemble the distribution of pre-non-linearity responses. Linear Nonlinear, in addition to approximately inverting the activation function, also approximates the activation function as an exponential function when re-applying it for the target model subject (step 3

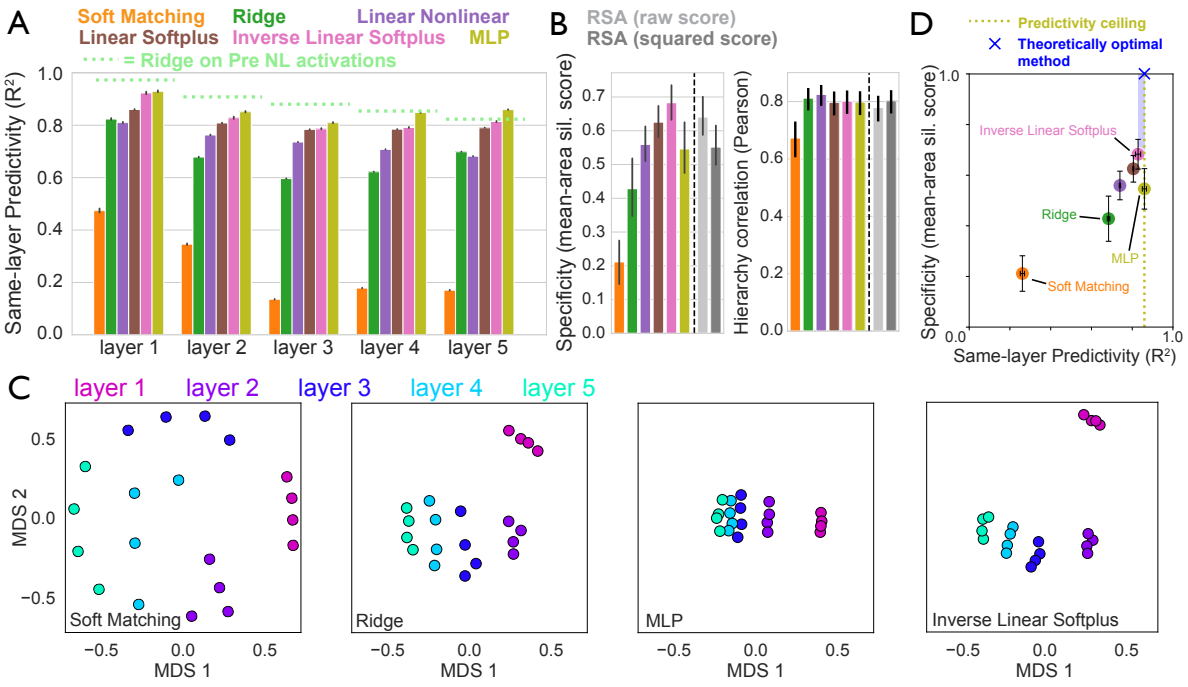

Figure 4: **Predictivity and specificity for a spectrum of candidate transform classes on a simulated model population.** (A) Same-layer predictivity when mapping responses between model subjects. (B) Specificity and hierarchy correlation for different transform classes. (C) Multidimensional scaling (MDS) plots to visualize dissimilarity scores, when mapping response profiles between all layers and all subjects. Each dot is a response profile for a particular subject and model layer. Distances between dots are optimized to match the dissimilarities using a particular comparison method. (D) A scatterplot comparing predictivity and specificity across transform classes. While the exact shape of the Pareto frontier for predictivity and specificity is unknown, we identify a bounded region (shaded blue) that contains at least one Pareto-optimal point. The diagonal of this region represents the maximum possible distance from our best IATC candidate (Inverse Linear Softplus) to the Pareto frontier.

of Fig. 3D). Even approximately accounting for the activation function improves predictivity compared to ridge regression, and the more precisely the activation function is accounted for, the better the predictivity (Fig. 4A) - that is, Inverse Linear Softplus outperforms Linear Softplus, which in turn outperforms Linear Nonlinear.

We next compare the predictivity of Inverse Linear Softplus to the maximum achievable same-layer predictivity, estimated using a 7-layer MLP trained on 1 million response patterns driven by ImageNet-train stimuli. The MLP does not yield substantially greater same-layer predictivity, providing some evidence that Inverse Linear Softplus is already close to the predictivity ceiling - as required for a viable IATC candidate.

Accounting for the activation function also improves area-identification specificity. A viable IATC candidate must not only align same-area responses between subjects, but also be as strict as possible, thus presumably able to separate responses from different layers. While Inverse Linear Softplus achieves high same-layer predictivity, it is not obvious that that it should also achieve high specificity. For example, if Inverse Linear Softplus improves predictivity by mapping more accurately between *any* pair of response profiles (including those from different layers), then we might see a decrease in specificity.

Our primary metric for specificity is a version of the silhouette score (Rousseeuw, 1987). A silhouette score close to 1 indicates that responses from different model layers (or brain areas) are well separated compared to responses from the same model layer (or brain area) (Fig. 2A). For a given response profile i, we compute:

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

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 response profiles from all other model layers. We take the mean score over all model subjects and layers. As an extension of specificity, we also look at the Pearson correlation between dissimilarity scores and distances between layers in the model hierarchy (Fig. 2B).

Inverse Linear Softplus increases specificity (Fig. 4B). The reason is that, by improving predictivity for the same-layer, Inverse Linear Softplus improves a key component of specificity, identification of same-layer similarity across subjects, while maintaining inter-layer separation (Fig. 4C).

Although classical RSA (Representational Similarity Analysis) is not a predictive mapping method (instead comparing summary statistics of population responses), we can still evaluate it for specificity as a useful benchmark against which our IATC candidates can be compared, as it is widely used for neural response comparisons (Kriegeskorte et al., 2008). We find that Inverse Linear Softplus outperforms RSA in terms of specificity (Fig. 4B).

Both very strict and very flexible methods impair specificity. Soft matching, a strict method that matches individual units between populations, is not only worse for predictivity (Fig. 4A), but also for specificity (Fig. 4B) as evaluated using the silhouette score. The reason is that soft matching (because of its low same-layer predictivity) has low same-layer identifiability and therefore low specificity (Fig. 4C). Soft matching also has lower specificity as evaluated using hierarchy correlation (Fig. 4B). This is because soft matching rates adjacent layers, such as layers 2 and 3, as dissimilar even compared to more distant layers, such as layers 2 and 4 (Fig. 4C).

At the other extreme, the flexible 7-layer MLP does not maximize specificity, because it reduces inter-layer separation, though the gain in same-layer identifiability slightly improves its specificity over ridge regression. This result illustrates the importance of identifying the strictest set of transforms that maps accurately across subjects. Inverse Linear Softplus and the MLP achieve similar levels of same-layer predictivity, but Inverse Linear Softplus is more constrained, leading to higher specificity.

These results illustrate the utility of the IATC, as attempting to identify it yielded a transform class that achieves high predictivity and high specificity. In fact, our best IATC candidate approaches the Pareto frontier for predictivity and specificity (Fig. 4D).

Testing IATC candidates for a mouse population

We now evaluate our methods on a mouse population, using Neuropixels recordings for 31 subjects in 6 brain areas averaged over 50 trials while the mice passively viewed 118 different visual stimuli. We evaluate methods for predictivity and specificity when comparing responses between mouse subjects and also examine their ability to separate candidate models of the mouse brain.

Rank order of transform classes is largely similar between mouse and model populations. The rank order of transform classes in terms of same-area predictivity provides evidence for which transform classes are better IATC candidates (despite the absolute scores being limited by relatively few response patterns for fitting the transforms, as well as a limited neuronal sample). As in the model population, soft matching achieves the lowest same-area predictivity, with linear regression performing substantially better (Figure 5A). Moreover, our biologically motivated transform classes that account for the activation function further improve predictivity on the mouse data over linear regression.

The fact that transform classes that account for the activation function are best for same-area predictivity hints at the possibility that, just as in the simulated population, the pre-non-linearity responses of two typical mouse subject are related by a linear transform, and the relationship between post-non-linearity responses is modified by the non-linearity. The fact that Linear Softplus (which models the activation function as a softplus) performs about the same as Linear Nonlinear (which models the activation function as an exponential) means that we cannot tell, based on our results, which of these activation functions better models the activation functions generating the mouse responses.

Strict methods that attempt to match individual units, such as soft matching, are worse for specificity compared to ridge regression and our biologically motivated transform classes (Figure 5B). This confirms that low predictivity can lead to low specificity by limiting same-area identifiability between subjects. Furthermore, this confirms that the most promising candidate IATCs - Linear Nonlinear and Linear Softplus - are constrained enough to separate responses from different brain areas, in addition to being flexible enough to align same-area responses between mouse subjects.

Bidirectional mapping can improve separation between brain models. We also evaluate how well each method separates different candidate models with respect to their assessed similarity to the mouse brain responses. We map 5 layers from four candidate models to the mouse responses: the ReLU-based AlexNet model of mouse visual cortex (Nayebi et al., 2022), our noisy softplus version of that model, a ResNet model trained on ImageNet categorization with 64x64 resolution stimuli, and a VGG-16 model trained on ImageNet categorization with 224x224 resolution stimuli (unlike the low resolution mouse visual system). We apply the same noise correction procedure for predictivity scores used in (Nayebi et al., 2022) to account for trial-to-trial variability (App. H). Model separation for a given area is evaluated as the absolute difference in assessed brain similarity between models (averaged over model pairs and model layers).

Typically, when mapping models to brains, model responses are mapped to brain responses, but the other direction of mapping is not considered. Guided by the IATC, we map bidirectionally, just as we do when aligning two mouse brains (Figure 5C). When mapping models to brain data, stricter mappings such as soft matching separate models more strongly, but the opposite pattern occurs when mapping brain data to models. When the scores for both mapping directions are averaged, the methods are all roughly comparable in terms of model separability. These results indicate that stricter methods like soft matching are not generally better for model separation. Furthermore, mapping in both directions can increase separation between models compared to unidirectional mapping from model to brain.

Overall, the results in this section show that our IATC results for the model population generalize to some extent to

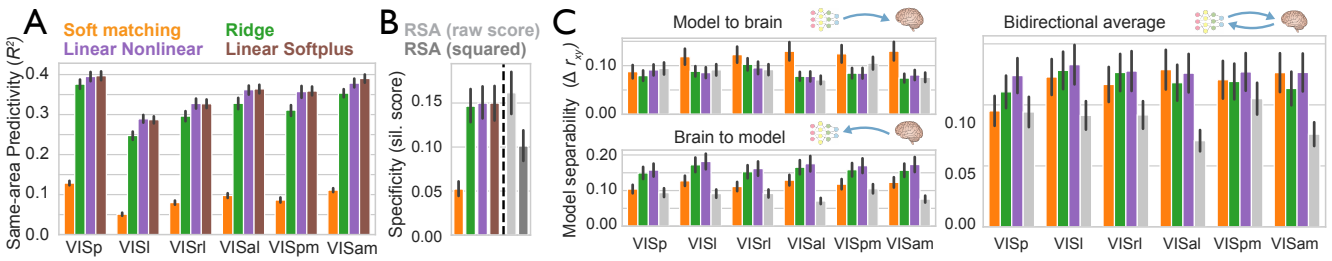

Figure 5: **Assessing candidate transform classes on a mouse population.** (A and B) Mapping responses from pooled mouse subjects to a held-out subject in order to evaluate same-area predictivity (A) and specificity (B). (C) We map five layers from four models to the mouse data: the ReLU-based AlexNet model of mouse cortex (Nayebi et al., 2022), our noisy softplus version of that model, ResNet, and VGG16. We evaluate the mean absolute difference between models (averaged over model pairs and model layers) in terms of assessed brain similarity (using the noise-corrected Pearson correlation or RSA score).

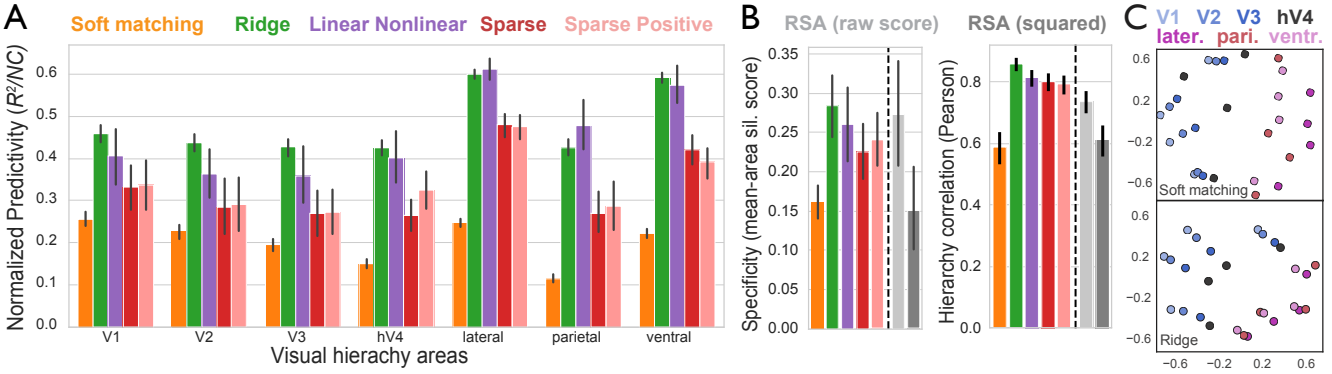

Figure 6: **Assessing candidate transform classes on a human population.** (A) Same-area predictivity on human subjects. We map each subject to each other subject for the same brain area, using fMRI responses to 1000 natural visual stimuli (Allen et al., 2021). We evaluate on 7 brain areas spanning different levels of the visual hierarchy. (B) Specificity and hierarchy correlation scores. To estimate distances in the visual hierarchy, we assign a hierarchy level of 1 to V1, 2 to V2, 3 to V3, 4 to hV4, and 5 to each of higher lateral, higher parietal and higher ventral. (C) MDS plots to visualize dissimilarity scores.

real brain data. In particular, we find evidence that the IATC for the mouse population is shaped by the non-linear activation function. Moreover, we again find that viable candidates for the IATC (such as Linear Nonlinear) are relatively good for both prediction and specificity. Our results also highlight the importance of the IATC's bidirectionality for model separation.

Testing IATC candidates on a human population

We evaluate transform classes on a large scale human fMRI dataset, the Natural Scenes Dataset (Allen et al., 2021), and evaluate methods for predictivity and specificity when mapping between human subjects for 7 visual areas: V1, V2, V3, hV4, as well as a higher area in each of the lateral, ventral, and parietal streams. Finally, we use IATC-guided bidirectional mapping to better separate between models of the human visual system.

Ridge regression achieves the best intra-area cross-subject predictivity. We again find that soft matching is unable to map across subjects with high predictivity. A more flexible transform, ridge regression, is needed to map more accurately across subjects (Figure 6A). Although Linear Nonlinear is close to ridge regression in terms of predictivity, it does not do noticeably better, perhaps because the low resolution of fMRI data obscures the effect of the non-linearity.

Ridge regression improves specificity and visual hierarchy identification. Under the mean-area silhouette score, soft matching performs worse than other methods while ridge regression performs best, suggesting that soft matching has lower specificity than ridge regression (Figure 6B). Ridge regression scores are more correlated with distances in the visual hierarchy, suggesting that ridge regression better tracks differences across the functional hierarchy. The improved hierarchical correlation for ridge regression compared to soft matching is apparent on an MDS plot visualizing distances between response profiles for different subjects and brain areas (Figure 6C).

We also consider sparse regressions (Prince et al., 2024), which use a lasso penalty to encourage sparse weights (with or without a positive weights constraint). These methods are stricter than ridge regression, but not as strict as soft

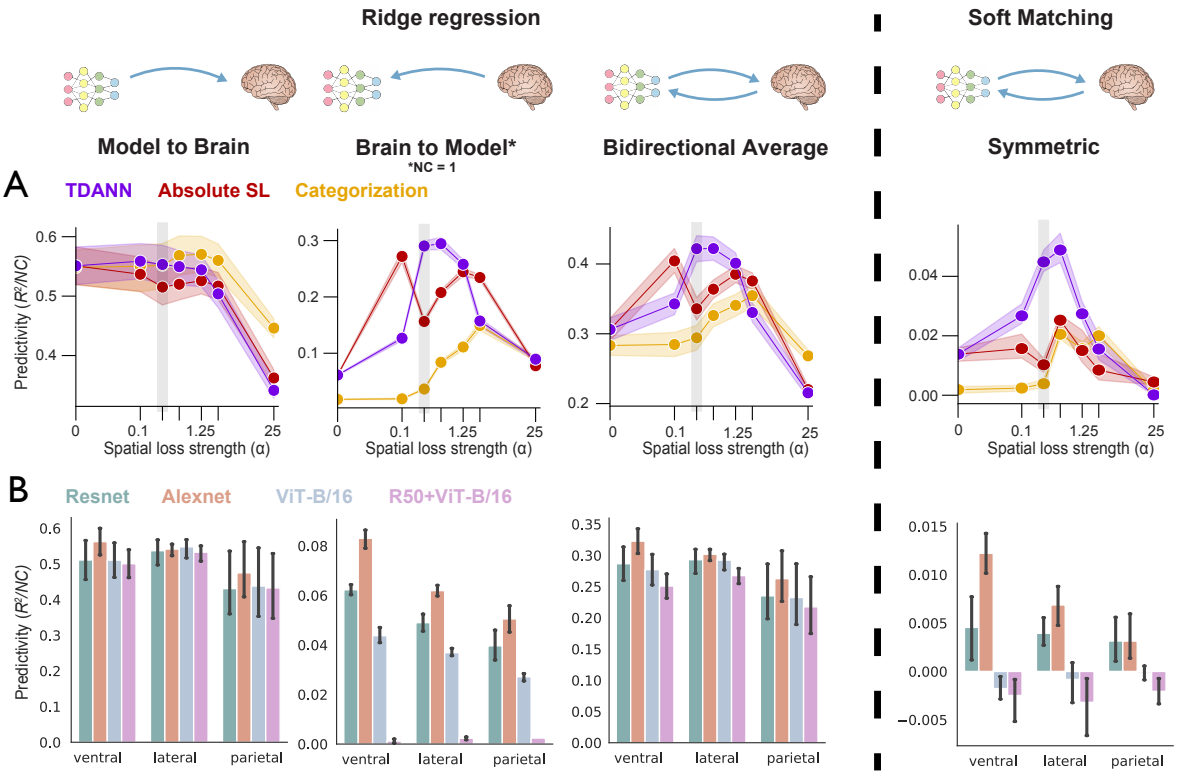

Figure 7: **Using bidirectional IATC-guided mapping to improve model separation.** We map between models and human fMRI responses (Natural Scenes Dataset), replicating analyses from Fig. 6A,B of Margalit et al. (2023) and Fig.4F of Khosla & Williams (2023) that compared linear regression (model to brain direction) to stricter methods that match individual units, such as 1-1 mapping or soft matching. Unlike the prior analyses, we map bidirectionally between models and brains, as required by the IATC approach. (A) Comparing topographic models (Margalit et al., 2023) with different training objectives (TDANN, Categorization, Absolute Spatial Loss), and spatial loss strengths (α) to higher ventral stream ROI. A key comparison is between the TDANN with intermediate spatial loss $\alpha = 0.25$ (highlighted with gray bar) and non-topographic models ($\alpha = 0$). The TDANN with $\alpha = 0.25$ was found in Margalit et al. (2023) to best match the brain based on a one-to-one mapping and based on predicting topographic organization of the visual cortex. (B) Comparing two CNN models (ResNet and Alexnet) and two transformer models (ViT-B/16 and R50+ViT-B/16) to higher visual areas, as in Khosla & Williams (2023).

matching. We observe a loss in predictivity and specificity for sparse regressions (Fig. 6A,B), highlighting that even somewhat stricter-than-linear methods can impair predictivity and specificity. Although we cannot compare ridge regression to a very flexible control such as an MLP given the dataset size (1000 response patterns), ridge regression seems to be the best IATC candidate.

Bidirectional IATC-guided mapping improves separation of candidate brain models. Recent work (Margalit et al., 2023; Finzi et al., 2022) introduced a topographic model (TDANNs) of the visual system that combines functional and spatial constraints. While the TDANN with an intermediate level of spatial loss strength $\alpha = 0.25$ predicted topographical properties of visual cortex better than alternative models, a key question has been whether the TDANN's response patterns quantitatively match neuronal responses better than non-topographic models. Here, the issue of a correct comparison method has been crucial, as unidirectional linear regression (from model to brain) failed to differentiate the TDANN from non-topographic models, while a 1-1 mapping did (Margalit et al., 2023, Fig. 6A,B). However, the very strictness of 1-1 mapping limited brain predictivity and the inter-animal noise ceiling, leaving it unclear how strong the evidence is in favor of the TDANN model. We therefore investigated whether IATC-guided methods could distinguish between the TDANN and alternative models in terms of matching neuronal responses.

Guided by the IATC, we did ridge regression in both directions between models and the brain. Although linearly mapping model responses to the brain data does not separate strongly between the models, linearly mapping the brain data to the model responses separates strongly between the models (Figure 7A), identifying the TDANN with $\alpha = 0.25$ or 0.5 as being the most brain-like, convergent with soft matching re-

sults and also with prior evidence in favor of that model (Margalit et al., 2023). In fact, model separation using bidirectional ridge regression is much larger than for soft matching. This result can be attributed to the fact that soft matching is an extremely strict method, which results in low predictivity scores for all models (Figure 7A, right-most plot). By distinguishing more strongly between TDANN ($\alpha = 0.25, 5$) and alternative models such as non-topographically constrained models ($\alpha = 0$), IATC-guided bidirectional ridge regression provides stronger evidence in favor of the TDANN.

Along similar lines, Khosla & Williams (2023) observed cases where linearly mapping models to brain responses did not separate models, but soft matching did. Revisiting this analysis but with bidirectional mappings, we found that bidirectional ridge regression increased model separation over soft matching (Fig. 7B), further confirming the utility of IATC-guided bidirectional mappings.

Discussion

The IATC provides a principled framework for model-brain comparison by identifying the strictest set of transforms needed to map neural responses accurately between animal subjects in a population (for the same brain area). We find in three settings (model population, mouse population, and human population) that a working estimate of the IATC achieves both high predictivity and high specificity, two key desiderata for a model-brain comparison method. In a simulated population of neural networks, we identified how the neuronal activation function shapes the IATC, leading to a transform class that improves predictivity and specificity relative to standard mapping classes. On a mouse electrophysiology dataset, we also find evidence that the IATC is constrained by the neuronal activation function, suggesting that IATC results for the model population can meaningfully generalize to real brain data. On a human dataset, the resolution of the fMRI data does not allow us to observe an effect of the activation function on the IATC, but we still see differentiation between candidate transform classes that is consistent with our findings from simulated population and mouse data. Moreover, we use the IATC-guided bidirectional mappings to enable better model-brain comparisons, uncovering new evidence differentiating topographic models of the visual system compared to non-topographic models.

When beginning this work, we assumed that the goals of specificity and predictivity would likely be in tension – with stricter methods (such as soft matching) being better for specificity of model identification and worse at prediction, and more flexible methods (such as linear regression) showing the opposite pattern (Ding et al., 2021; Kornblith et al., 2019; Conwell et al., 2022; Finzi et al., 2022). However, the intuition that stricter methods are generally better for specificity overlooks the fact that specificity requires identifying high similarity across subjects for responses of the same type, not just separating responses of different types. Extremely strict methods fail to align same-area responses well across subjects, leading to low identifiability and thus low specificity (Fig. 2A). On each population, the best working estimate of the IATC improves same-area identifiability by mapping responses accurately across subjects, while still maintaining inter-area separation, leading to high specificity. Thus, there is, in fact, no real tension between specificity and predictivity.

A key aspect of the IATC approach is to map bidirectionally between models and brains, not just in the model to brain direction, just as when comparing two brains to each other. Considering both directions can reveal cases where a given model contains spurious features, improving model separation. Unlike previous works that motivated symmetry with the assumption that similarity scores must be distance metrics (Williams et al., 2021; Khosla & Williams, 2023), under our IATC approach, similarity naturally emerges from bidirectional relationships between brains in a population. Our approach also differs from work that treated both directions of mapping as separate, potentially inconsistent methods (Soni et al., 2024) rather than as two components of a single method. In future work, we plan to further investigate bidirectional mappings and address potential limitations, such as how subsampling of neurons may affect the brain-to-model direction.

A limitation of our results using a simulated population is the question of whether the sources of variability we consider (different seeds for weight initialization and training data order) are anything like sources of actual brain variation. In future work, we will work towards a "generative model" that more accurately describes inter-subject variability.

A second direction for improvement will be to improve IATC estimation. Rather than evaluating a small set of transform classes as done in this paper, we hope to systematically learn the IATC in a data-driven fashion using large-scale optimization techniques. An especially exciting possibility will be to use such data-driven estimation methods to strengthen and refine the preliminary results we show here suggesting that neural circuit mechanisms constrain the IATC. This will be an increasingly realistic prospect as neural datasets grow in size (with many subjects, neurons per subject, and stimuli).

Recent work has shown that neural networks, and likely the brain, have privileged axes in their representations: unit-level tuning curves that are similar between subjects, and which cannot be linearly remixed without constraint (Khosla & Williams, 2023). A natural synthesis of this result with our findings would be a hybrid "Linear Subspace-Nonlinear" transform class, similar to the Linear Nonlinear transform class that models the activation function, but where the linear portion of the transform is constrained by the preferred axes to specific subspaces of allowable transforms. This hybrid would be stricter than full Linear-Nonlinear, but more flexible than soft matching. We hypothesize that such a transform class would be a better estimate of the true IATC and would occupy a "sweet spot" on the strictness-flexibility continuum (Fig. 8). Actually estimating these transform subspaces will likely involve novel data-driven discovery and optimization methods.

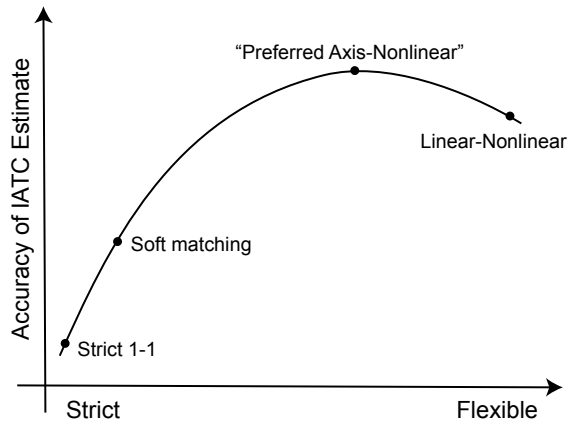

Figure 8: **Taking preferred axes into account when estimating the IATC.** A hypothetical "Preferred Axis-Nonlinear" transform class could be a more accurate IATC estimate.

Acknowledgements

This work was supported by the following awards: To I.T.: Patrick Suppes Philosophy of Science Dissertation Award. To D.L.K.Y.: Simons Foundation grant 543061, National Science Foundation CAREER grant 1844724, National Science Foundation Grant NCS-FR 2123963, Office of Naval Research grant S5122, ONR MURI 00010802, ONR MURI S5847, and ONR MURI 1141386 - 493027. We also thank the Stanford HAI, Stanford Data Sciences and the Marlowe team, and the Google TPU Research Cloud team for computing support.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from `https://www.tensorflow.org/` (Software available from tensorflow.org)

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... Kay, K. (2021, December). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. Retrieved from `http://dx.doi.org/10.1038/s41593-021-00962-x` doi: 10.1038/s41593-021-00962-x

Canatar, A., Feather, J., Wakhloo, A., & Chung, S. (2024). A spectral theory of neural prediction and alignment. *Advances in Neural Information Processing Systems*, *36*.

Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 1–taking mechanistic abstraction seriously. *arXiv preprint arXiv:2104.01490*.

Chichilnisky, E. (2001). A simple white noise analysis of neuronal lightresponses. *Network: computation in neural systems*, *12*(2), 199.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, 2022–03.

Ding, F., Denain, J.-S., & Steinhardt, J. (2021). Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, *34*, 1556–1568.

Feather, J., Leclerc, G., Mkadry, A., & McDermott, J. H. (2023). Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, *26*(11), 2017–2034.

Finzi, D., Margalit, E., Kay, K., Yamins, D. L., & Grill-Spector, K. (2022). Topographic dcnns trained on a single self-supervised task capture the functional organization of cortex into visual processing streams. In *Svrhm 2022 workshop@ neurips.*

Gao, X., Wen, M., Sun, M., & Rossion, B. (2022). A genuine interindividual variability in number and anatomical localization of face-selective regions in the human brain. *Cerebral Cortex*, *32*(21), 4834–4856.

Haug, H. (1987). Brain sizes, surfaces, and neuronal sizes of the cortex cerebri: a stereological investigation of man and his variability and a comparison with some mammals (primates, whales, marsupials, insectivores, and one elephant). *American Journal of Anatomy*, *180*(2), 126–142.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cor-

tical representation. *PLoS computational biology*, *10*(11), e1003915.

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

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. In *International conference on machine learning* (pp. 3519–3529).

Kriegeskorte, N., & Diedrichsen, J. (2016). Inferring brain-computational mechanisms with models of activity measurements. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1705), 20160278.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 4.

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., & Yamins, D. L. (2023). A unifying principle for the functional organization of visual cortex. *bioRxiv*.

McCullagh, P., & Nelder, J. (2019). *Generalized linear models*. Routledge.

Mineault, P., Bakhtiari, S., Richards, B., & Pack, C. (2021). Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *Advances in Neural Information Processing Systems*, *34*, 28757–28771.

Nayebi, A., Attinger, A., Campbell, M., Hardcastle, K., Low, I., Mallory, C. S., ... others (2021). Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. *Advances in Neural Information Processing Systems*, *34*, 12167–12179.

Nayebi, A., Kong, N. C., Zhuang, C., Gardner, J. L., Norcia, A. M., & Yamins, D. L. (2022). Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *bioRxiv*, 2021–06.

Nayebi, A., Rajalingham, R., Jazayeri, M., & Yang, G. R. (2024). Neural foundations of mental simulation: Future prediction of latent representations on dynamic scenes. *Advances in Neural Information Processing Systems*, *36*.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Prince, J. S., Conwell, C., Alvarez, G. A., & Konkle, T. (2024). A case for sparse positive alignment of neural systems. In *Iclr 2024 workshop on representational alignment.*

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., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Soni, A., Srivastava, S., Kording, K., & Khosla, M. (2024). Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. *bioRxiv*, 2024–08.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of cognitive neuroscience*, *33*(10), 2044–2064.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., ... others (2023). Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature neuroscience*, *18*(7), 1025–1033.

Thompson, B., Tilly, J., dependabot[bot], Santorella, E., Schmidt, M.-A., Bittarello, L., ... pwojtasz (2025). *Quantco/glum*. Retrieved from `https://github.com/Quantco/glum`

Van Vreeswijk, C., & Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, *274*(5293), 1724–1726.

Wang, P. Y., Sun, Y., Axel, R., Abbott, L., & Yang, G. R. (2021). Evolving the olfactory system with machine learning. *Neuron*, *109*(23), 3879–3892.

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.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118.

Zipser, D., & Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*(6158), 679–684.

Appendix

A Central hypercolumn selection

In order to map between units with similar functional roles (at least for the same model layer), we do our model-model fits using only the central hyper-column of units in each layer (i.e. the units whose receptive field is directly at the middle of the input image). Indeed, even when constraining the mapping to use only the central hyper-column, we are able to identify high similarity across model instances for the same layer, at least when assessing pre-non-linearity responses using a linear transform.

B Soft matching as a transform class

While Khosla & Williams (2023) do not explicitly formulate the soft matching score as a predictive mapping, it can be formulated as one. Computing the soft matching score involves maximizing:

$$\Sigma_{i,j}\mathbf{T}_{ij}\mathbf{C}_{ij}$$

where $\mathbf{T}$ is the transport matrix, subject to the constraints that the columns of the matrix sum to $1/N_Y$, while the rows sum to $1/N_X$, and $\mathbf{C}$ is the matrix of Pearson correlations between each source neuron and each target neuron. The transport matrix can be interpreted as a joint probability distribution over source neurons and target neurons (where the marginal distributions are uniform discrete). Thus, the soft matching score is the *expected* correlation between source and target neurons, according to joint probabilities encoded by the optimal transport matrix.

Since maximizing the above objective requires identifying source neurons that are highly correlated with each target neuron, we can use the source neurons to predict the value of each target neuron, weighted by the probabilities in the optimal transport matrix. First, for each source neuron X_i and target neuron Y_j, we can predict Y_j's responses across a set of stimuli (symbolized as the vector $\mathbf{Y}_j$) based on X_i's responses to those stimuli (symbolized as $\mathbf{X}_j$) as:

$$\hat{\mathbf{Y}}_j = \frac{\sigma(\mathbf{Y}_j)}{\sigma(\mathbf{X}_i)}[\mathbf{X}_i - \bar{\mathbf{X}}_i]\mathbf{C}_{ij} + \bar{\mathbf{Y}}_j$$

This is essentially using the correlation $\mathbf{C}_{ij}$ to do ordinary least squares between X_i's responses and Y_j's responses.

For a single target neuron Y_j, we compute the expected value of these correlation-based predictions across source neurons, if we sampled source neurons according to the conditional probability distribution $P(X = X_i|Y = Y_j)$. Since $\mathbf{T}_{ij} = P(X_i, Y_j)$ and $P(Y_j) = 1/N_Y$, it follows that $P(X = X_i|Y = Y_j) = N_Y\mathbf{T}_{ij}$. Using these conditional probabilities, the overall prediction $\hat{\mathbf{Y}}_j$ then becomes:

$$\hat{\mathbf{Y}}_j = N_Y\sigma(\mathbf{Y}_j)\Sigma_i\frac{\mathbf{X}_i - \bar{\mathbf{X}}_i}{\sigma(\mathbf{X}_i)}\mathbf{T}_{ij}\mathbf{C}_{ij} + \bar{\mathbf{Y}}_j$$

C Motivating the softplus activation function with a simple model of a noisy spiking process

Our simulation of spike counts is based on the following highly simplified model. We assume that a neuron receives total input $X \sim \mathcal{N}(\mu, \sigma^2)$ and that during a single time interval equal to the neuron's refractory period (which we assume to be about 1 ms), the neuron either fires once or not at all, depending on whether $X > T$, where T is a fixed threshold (Fig. 9A). We count the number of spikes over a 100 ms time range, and average over 100 trials.

Under this model, the total mean (over trials) spike count $S_t(\mu)$ over a time period t (expressed as a function of the mean total input to the neuron μ) is equal to $t/R * \Phi(\mu - T, \sigma^2)$, where Φ is the Gaussian CDF. This means that the activation function should have a sigmoid shape, which saturates at sufficiently high mean inputs (Fig. 9B). However, many cortical neurons are thought to fire in the fluctuation driven, unsaturated regime (Van Vreeswijk & Sompolinsky, 1996). We therefore focus on unsaturating functions like softplus and fit these functions to spike counts that we simulated in the unsaturated regime (Fig. 9C). The softplus function is defined as:

$$\text{softplus}(x) = \ln(1 + e^x)$$

D Noisy Softplus AlexNet models

To obtain our noisy softplus variant of the AlexNet mouse model, every ReLU sub-layer in the AlexNet models is exchanged for a Softplus sub-layer followed by a Poisson-like noise block whose mean is the output of the Softplus sub-layer. PyTorch enables noisy models to be trained using a reparameterization trick, but only for certain probability distributions (not for the Poisson distribution). We use the Gamma distribution as a stand-in for Poisson, choosing shape parameter $k = \lambda$, where λ is the Poisson parameter (which is chosen to be the output of the Softplus sub-layer), and scale parameter $\theta = 1$. This allows us to replicate two statistical properties of Poisson variables: non-negative samples and variance-mean ratio of 1, both of which are important for using Linear Nonlinear, Linear Softplus and Inverse Linear Softplus (which all use a Poisson GLM) to predict the responses.

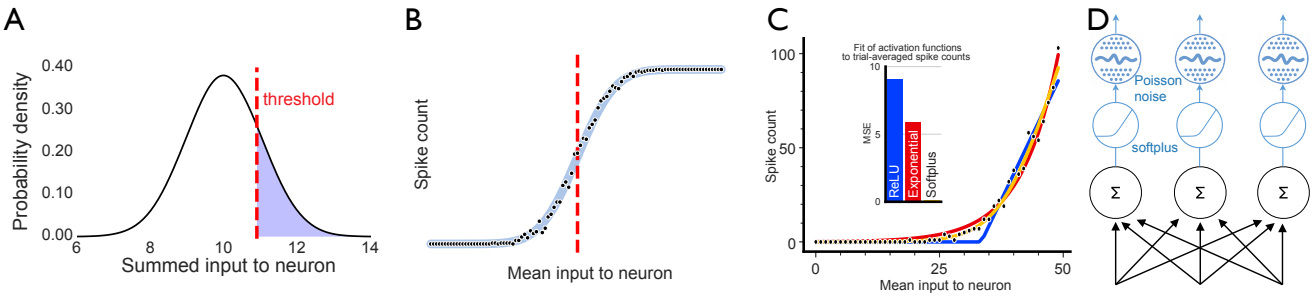

Figure 9: **A more biologically consistent activation function.** (A) Biological activation functions are the result of a noisy spiking process. Because summed inputs to neurons are noisy, the firing probability is positive even when the mean input is sub-threshold. Here, the probability of spiking is represented as the size of the blue region. (B) The resulting activation function, unlike ReLU, is strictly positive and increasing. Dots represent simulated spike counts, which are Poisson-distributed in the limit of very small firing rates. (C) Fitting different activations to simulated spike counts, allowing for scaling and translation. Softplus fits spike counts the best in the sub-threshold regime. The exponential activation also function performs somewhat better than ReLU. Intuitively, the reason ReLU does not fit as well is that it has a hinge that prevents it from capturing the smooth increase in firing rate. Spike counts are plotted for a single trial. (D) We replaced each ReLU non-linearity in the models with a softplus non-linearity and a Poisson-like noise sampler.

To avoid numerical difficulties for small values of $k = \lambda$, we scale the softplus outputs by 100 before sampling from the Gamma distribution. We then train the noisy softplus models so that their instance recognition training score (as well as validation score on ImageNet categorization) are equal to those of the ReLU-based AlexNet models.

E Inverting the softplus function in Inverse Linear Softplus

In order to invert the softplus non-linearity as the first step of Inverse Linear Softplus, we apply the inverse of the softplus function to the softplus model responses in a given layer, averaged over 50 trials. Because the softplus outputs at every model layer are scaled by 100 before taking Poisson-like samples from the Gamma distribution (App. D), we un-scale the trial-averaged responses before applying the softplus inverse. The inverse of the softplus function is well-defined (because softplus is strictly increasing) and has the following formula:

$$\text{softplus}^{-1}(y) = \ln(e^y - 1)$$

In practice, to avoid numerical difficulties for very small values of y, we do not apply this formula directly and instead use a more numerically stable implementation of the softplus inverse adapted from the TensorFlow library (Abadi et al., 2015).

F Yeo-Johnson scaling in Linear Nonlinear and Linear Softplus

When the activation function is known exactly and its inverse is well-defined (as in the case of the Softplus-based model), we can directly invert the activation function to recover the pre-non-linearity responses. However, when mapping animals to animals (or, if enough neurons are measured, animals to models), we cannot easily invert the activation function if we do not know its exact form for a given neuron. Yeo-Johnson scaling uses a power transformation to make the features closer to normally distributed over the stimuli. We expect this transformation to make the post-non-linearity features more correlated with pre-non-linearity responses because the non-linear activation function skews the distribution of the pre-non-linearity responses (which are roughly normally distributed over the stimuli). Indeed, we find that Yeo-Johnson scaling noticeably increases the Pearson correlation (Fig. 10) with the pre-non-linearity responses for the noisy softplus models, almost as much as if you had directly applied the inverse of the softplus activation function to the post-non-linearity responses. We hypothesize that Yeo-Johnson scaling has a similar effect in the case of animal firing rates.

We implement Yeo-Johnson scaling with the PowerTransformer class in sklearn (Pedregosa et al., 2011). The power transform fits one parameter. To implement Yeo-Johnson scaling as the first step of Linear Nonlinear or Linear Softplus, we put the PowerTransformer object followed by a GLM object into an sklearn Pipeline, so that the power parameter is only fit on the training data, not on test data.

G Implementation details of GLMs

The GLM object is created using the *glum* package (Thompson et al., 2025). Each GLM specifies the inverse link function that relates the linear prediction to the response variable (such as ReLU, exponential or softplus), and the assumed noise structure

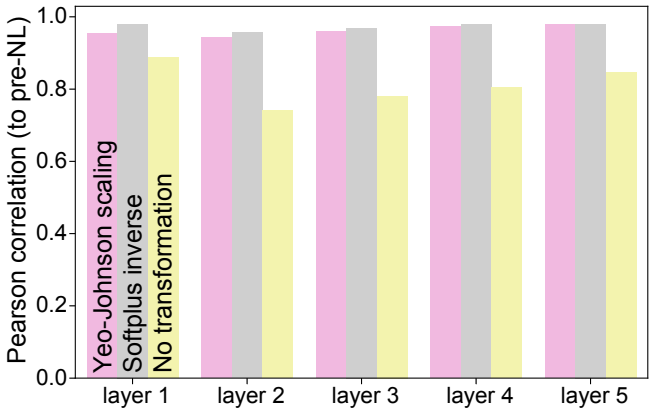

Figure 10: **Correlation between post-non-linearity responses and pre-non-linearity responses after transforming the post-non-linearity responses in different ways (responses are for the noisy softplus models, averaged over 50 trials).** We focus on correlation here because Yeo-Johnson scaling does not improve the R^2 score with respect to pre-non-linearity features (i.e. it does not directly match them), which makes sense as it is merely unskewing the distribution of post-NL features, which are already rather correlated with pre-NL features. Nevertheless, increased correlation implies that the pre-NL features can be more easily matched after linear re-weighting, as is done in Linear Nonlinear or Linear Softplus.

in the response variable (Poisson noise in the case of Linear Nonlinear or Linear Softplus). The weights of the specified GLM are then optimized through Iterative Reweighted Least Squares.

The inverse link function in Linear Softplus or Inverse Linear Softplus involves a scaling parameter c:

$$\hat{y} = c * \text{softplus}(\theta^T x)$$

where $\hat{y}$ is the output of the inverse link function (i.e. the predicted values for the target responses), x is the vector of predictors (trial averaged responses of the source model after applying either Yeo-Johnson scaling or exactly inverting the activation function) and θ is the fitted linear weights. When predicting noisy softplus model responses, we set $c = 100$, the same softplus output scaling we used when training the models themselves. But when fitting Linear Softplus to predict mouse responses, we do not know *a priori* the optimal scaling parameter and must cross-validate values of c along with the ridge penalty using GridSearchCV in sklearn.

H Noise correction when comparing models to mouse data

When mapping between model responses and trial-averaged mouse responses (Fig. 5C), it is important to account for trial-to-trial variability. Here we briefly describe the noise correction procedure that is used to obtain more accurate predictivity scores. The full derivation of this procedure is found in Nayebi et al. (2022). The goal of the procedure is to accurately estimate the following quantity:

$$\text{Corr}(\mathcal{M}(r_{\text{train}}; t^B_{\text{train}})_{\text{test}}, t^B_{\text{test}})$$

where $\mathcal{M}$ is a given mapping method (such as ridge regression), r_{train} is the model responses in a given layer (which, except for the noisy softplus models, are deterministic) over the training stimuli, t^B_{train} is the *true* trial-averaged (averaged over the ideal limit of infinitely many trials) responses of a particular subject and brain area B over training stimuli, $\mathcal{M}(r_{\text{train}}; t^B_{\text{train}})_{\text{test}}$ are the test predictions under the mapping method of the target animal's responses over the test stimuli using the model responses as predictors, and t^B_{test} are the actual ground-truth responses of the target animal over test stimuli. This quantity cannot be directly computed because we do not have infinitely many trials per stimulus, and instead must estimate it based on finitely many trials (50 trials per stimulus in the case of the Allen Institute mouse data).

To perform the noise correction, we use bootstrapping. For each bootstrapped sample, we separate the $N = 50$ trials into two split halves of 25 trials each (indexed in the notation given below by 1 and 2) and take the trial-averaged response for each stimulus for each split half of those trials. Then the noise-corrected predictivity (in terms of Pearson correlation) is computed as:

$$\text{median}\left\langle \frac{\text{Corr}\left(\mathcal{M}(r^{\ell}_{\text{train}}; s^B_{1,\text{train}})_{\text{test}}, s^B_{2,\text{test}}\right)}{\sqrt{\widetilde{\text{Corr}}\left(\mathcal{M}(r^{\ell}_{\text{train}}; s^B_{1,\text{train}})_{\text{test}}, \mathcal{M}(r^{\ell}_{\text{train}}; s^B_{2,\text{train}})_{\text{test}}\right) \times \widetilde{\text{Corr}}\left(s^B_{1,\text{test}}, s^B_{2,\text{test}}\right)}} \right\rangle$$

where the median is computed over the target animal's neurons, and the $\langle ... \rangle$ represents an average over all bootstrap samples. The $\widetilde{\text{Corr}}$ represents a Spearman-Brown corrected Pearson correlation rather than a raw Pearson correlation. $s^B_{\text{i, train/test}}$ represents the trial-averaged responses of the subject for split-half i (which is either 1 or 2) over the train stimuli or test stimuli (unlike t which was the ideal trial-average over infinitely many trials).

In most cases, we use 100 bootstrapping samples, and 10 train-test splits. However, in some cases, we use fewer bootstrapping samples and train-test splits because of computation time constraints. In particular, whenever we map from VGG-16 to mouse responses, or whenever we use Linear Nonlinear, we use 16 bootstrapping samples and 1 train-test split.

I Noise correction when comparing models to human fMRI data

The bootstrapping approach to noise correction described in App. H is not possible in the case of the human fMRI data, where there are only 3 trials per stimulus, not 50 trials. We instead use the method of noise correction recommended in Allen et al. (2021). The idea is to simply divide the raw R^2 predictivity (with respect to a given target voxel) by the noise ceiling of that target voxel, which is computed as:

$$NC = \frac{\text{ncsnr}^2}{\text{ncsnr}^2 + \frac{1}{n}}$$

where ncsnr stands for "noise ceiling signal-to-noise ratio" and is provided for each voxel with the Natural Scenes Dataset, and n is the number of trials (3). This accounts for trial-to-trial variability when mapping model units to noisy voxel responses. When mapping in the other direction, from noisy voxels to model units, we set $NC = 1$, since the models of the human visual system we consider are all deterministic. It is worth noting that this procedure can only account for noise in the target voxel, but cannot account for noise in the source predictors (which is certainly an issue when either mapping brain-to-brain or brain-to-model). Developing statistical methods to correct for source noise is a major open challenge for future research.