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

2 Artificial neural network models have emerged as promising mechanistic models of brain function, but there is 3 4 little consensus on the correct method for comparing 5 activation patterns in these models to brain responses. Drawing on recent work on mechanistic models in phi-6 losophy of neuroscience, we propose that a good com-7 8 parison method should mimic the Inter-Animal Transform Class (IATC) - the strictest set of functions needed to 9 accurately map neural responses between subjects in a 10 population for the same brain area. Using the IATC, we 11 can map bidirectionally between model responses and 12 brain data, assessing how well the model can masquer-13 ade as a typical subject using the same kinds of trans-14 forms needed to map across animal subjects. We attempt 15 to empirically identify the IATC in three settings: a simu-16 lated population of neural network models, a population 17 of mouse subjects, and a population of human subjects. 18 In each setting, we find that the empirically identified IATC 19 enables accurate neural predictions while also achiev-20 ing high specificity (i.e. distinguishing response patterns 21 from different areas while strongly aligning same-area re-22 sponses between subjects). In some settings, we find 23 evidence that the IATC is shaped by specific aspects of 24 the neural mechanism, such as the non-linear activation 25 26 function. Using IATC-guided transforms, we obtain new evidence, convergent with previous findings, in favor of 27 topographical deep neural networks (TDANNs) as models 42 28 of the visual system. 29

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

32

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; KhalighRazavi & Kriegeskorte, 2014; Sussillo et al., 2015; Wang et
al., 2021; Schrimpf et al., 2021; Mineault et al., 2021; Nayebi



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.

40 et al., 2021, 2024). These results naturally raise the question
41 of whether these models can serve as mechanistic models
42 of brain function, at least at some level of abstraction (Cao &
43 Yamins, 2021; Kriegeskorte & Diedrichsen, 2016). Although a
44 variety of methods have been proposed for quantitatively as45 sessing neural response similarity between models and brains
46 (Sucholutsky et al., 2023), there is little consensus on what the
47 correct method is.

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, ⁵⁵ inter-subject variability can be substantial – in humans, the
⁵⁶ estimated number of neo-cortical neurons can vary between
⁵⁷ subjects by up to a factor of 2 (Haug, 1987), and even the ex⁵⁸ act number of functionally identified brain areas can vary (Gao
⁵⁹ et al., 2022). As a result, 1-1 matching is likely to be problem⁶⁰ atic. A more sophisticated approach to model-brain mapping
⁶¹ is therefore needed.

A first generation of approaches to this problem used lin-62 ear mappings to compare models to brains (Yamins et al., 63 2014), but concerns have arisen that the linear mapping class 64 is too flexible to strongly separate models of the brain (Ko-65 rnblith et al., 2019; Ding et al., 2021; Conwell et al., 2022). 66 More recent methods have thus focused on "stricter" mapping 67 68 classes, such as soft matching (Khosla & Williams, 2023), that tighten the criteria for model-brain similarity while still allowing 69 comparisons between different-sized populations of neurons 70 (unlike 1-1 matching). 71

Here we develop the idea of the Inter-Animal Transform 72 Class (IATC), a concept that has been introduced in philoso-73 phy of neuroscience to handle the problem of between-subject 74 variability when building mechanistic models (Cao & Yamins, 75 2021). As defined in that work, the IATC is the strictest (small-76 est) class of functions that maps responses between any two 77 subjects in a natural population, matching the same brain area 78 across subjects with as high accuracy as possible (Fig. 1B).¹ 79 Both the strictness and mapping accuracy criteria are impor-80 tant: a good IATC candidate must (by virtue of mapping accu-81 racy) succeed in aligning same-area responses across sub-82 jects, while (by virtue of strictness) separating responses from 83 different areas. This pair of desiderata naturally suggests a 84 85 meta-metric of *specificity*, which evaluates the extent to which mapping simultaneously achieves high cross-subject within-86 area identifiability while maintaining between-area separabil-87 ity (Fig. 2A). As an extension of specificity, we also consider ¹¹¹ 88 whether similarity scores under a mapping method correlate 89 with inter-area distances in a known hierarchy (Fig. 2B). 90

Though defined relative to a population of real individuals, 91 the IATC can be used to compare artificial networks to brains. 92 Specifically, we propose to use the empirically-identified IATC 93 itself to map between models and brains - in effect, mea-94 suring how well the model can masquerade as a member of 95 the population (Fig. 1C). A key implication of this IATC-based 96 approach is that model-brain mappings should be performed 97 bidirectionally between models and brain data, just as when 98 comparing two brains to each other, rather than only mapping 99 one direction (from model to brain). 100 in

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



Figure 2: Evaluating candidate transform classes for specificity and hierarchy. (A) It is desirable for a mapping to achieve high *specificity* – simultaneously achieving withinarea *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.

107 methods, such as linear regression and soft matching, against
108 the two basic IATC criteria: they should map responses across
109 subjects within an area as accurately as possible, while sepa110 rating responses from different areas.

We first evaluate candidate transform classes on a simulated population of artificial neural network models. Because 112 we have complete knowledge of the network structure and can 113 "measure" responses for all units over many stimuli, we are 114 able to observe how the specific form of the activation func-115 tion in the network shapes the relationships between model 116 subjects' responses. This motivates a new transform class 117 that maps responses across model subjects with close-to-118 maximum predictivity and high specificity, yielding a reason-119 able estimate of the IATC. We then evaluate these different 120 methods on real neural datasets, including both mouse elec-121 122 trophysiology and human fMRI recordings.

Results

124 Testing candidate IATCs for a simulated population

123

¹²⁵ We first evaluate transform classes on a simulated popula-¹²⁶ tion of neural networks (Figure 3A) against the IATC crite-¹²⁷ ria by testing for same-area predictivity as well as for speci-¹²⁸ ficity. Our simulated population consists of neural networks ¹²⁹ based on a state-of-the-art model of mouse visual cortex: ¹³⁰ an AlexNet trained with contrastive learning on 64x64 inputs ¹³¹ (Nayebi et al., 2022). We further modified the model to use

¹The maximum possible accuracy that is obtainable may be less ¹²⁹ than perfect because different subjects' neural representations can ¹³⁰ have different metamers (Feather et al., 2023), and therefore different ¹³¹ neural encoding functions with different null spaces.



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 nonlinearity. 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 postnon-linearity activations.

а 132 133 ate a population of model subjects, we vary the random seed 167 IATC candidate that works for post-non-linearity responses. 134 ontrolling the weight initialization and training data order. We 135 map responses between subjects using 10000 activation pat-136 terns driven by ImageNet-validation stimuli (80/20 train-test 137 split), evaluating the test R^2 , median across target neurons for 138 given model layer, averaged in both directions and across 139 а all pairs of subjects. 140

141 142 143 144 145 146 147 148 149 150 as those implemented by an MLP, which if true would suggest 184 (Fig. 4A). 151 that model subjects trained from different random seeds are 185 152 highly dissimilar in their learned representations. 153

154 155 156 157 158 159 160 161 162 163 164

softplus activation function followed by Poisson-like noise 165 thought of as corresponding to trial-averaged firing rates (Figto better mimic neuronal response characteristics. To gener- 166 ure 3C). Because EPSPs are hard to measure, we develop an

168 Improving cross-subject mapping by considering the 169 non-linear activation function. The results for ridge re-170 gression pre- and post-non-linearity suggest that an ideal IATC candidate must consider the effect of the non-linearity on 171 the relationships between different subjects' post-non-linearity 172 173 responses. We develop a transform class called Inverse Lin-The activation function has a substantial effect on 174 ear Softplus, which inverts the softplus activation function, apsame-layer response similarity between model subjects. 175 plies a fitted linear mapping between the two subjects, and Because a viable IATC candidate must map responses accu- 176 re-applies the softplus activation to predict the target subject's rately across different subjects for the same layer, we first eval- 177 post-softplus responses (Fig. 3D). Building on an established uate ridge regression, which has been widely used for neural 178 framework of generalized linear models (GLMs) (McCullagh & esponse prediction (Canatar et al., 2024). Surprisingly, ridge 179 Nelder, 2019; Chichilnisky, 2001), we use a GLM whose inegression achieves only moderate same-layer predictivity for 180 verse link function is precisely matched to the softplus activaintermediate model layers when mapping post-softplus acti- 181 tion in order to fit the linear mapping and re-apply the softplus vations between subjects (Fig. 3B). This raises the possibility 182 activation function. This yields substantially higher predictivity that the IATC might require highly non-linear transforms such 183 than ridge regression when mapping post-softplus activations

In the case of real neural data, we may not know the ex-186 act form of the activation function, and we therefore develop However, we observe a "zippering" effect: at each layer, 187 versions of our transform class that attempt to approximately pre-non-linearity responses are close to linearly related be- 188 account for the activation function. Linear Softplus (unlike tween subjects, but the activation function disrupts these lin- 189 Inverse Linear Softplus) approximately inverts the activation ear relationships for post-non-linearity responses, before the 190 function (step 1 of Fig. 3D) for the source model subject by next layer's pre-non-linearity responses become linearly re- 191 using Yeo-Johnson scaling. Yeo-Johnson scaling applies a lated again (Figure 3B). This effect suggests that the model 192 power transformation to make the post-non-linearity features subjects are actually similar, despite the apparent divergence 193 normally distributed, and thus more closely resemble the dissuggested by the failure of ridge regression to map post-non- 194 tribution of pre-non-linearity responses. Linear Nonlinear, in linearity responses accurately. Pre-non-linearity activations 195 addition to approximately inverting the activation function, also can be thought of as corresponding to trial-averaged EPSPs 196 approximates the activation function as an exponential funcin real neurons, while post-non-linearity activations can be 197 tion 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.

198 function improves predictivity compared to ridge regression, 221 199 and the more precisely the activation function is accounted 200 for, the better the predictivity (Fig. 4A) - that is, Inverse Lin-201 ear Softplus outperforms Linear Softplus, which in turn out-202 performs Linear Nonlinear. 203

We next compare the predictivity of Inverse Linear Softplus 204 to the maximum achievable same-layer predictivity, estimated 205 using a 7-layer MLP trained on 1 million response patterns 206 driven by ImageNet-train stimuli. The MLP does not yield sub-207 208 stantially greater same-layer predictivity, providing some evidence that Inverse Linear Softplus is already close to the pre- 222 where a(i) is the mean dissimilarity between i and other re-209 dictivity ceiling - as required for a viable IATC candidate. 210

Accounting for the activation function also improves 211 area-identification specificity. A viable IATC candidate 212 must not only align same-area responses between subjects, 213 but also be as strict as possible, thus presumably able to sep-214 arate responses from different layers. While Inverse Linear 215 Softplus achieves high same-layer predictivity, it is not obvious 216 that that it should also achieve high specificity. For example, if 217 Inverse Linear Softplus improves predictivity by mapping more 218 accurately between any pair of response profiles (including 219

of Fig. 3D). Even approximately accounting for the activation 220 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:

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

sponse profiles for the same model layer, and b(i) is the mean dissimilarity between *i* and response profiles from all other 224 model layers. We take the mean score over all model sub-225 226 jects and layers. As an extension of specificity, we also look 227 at the Pearson correlation between dissimilarity scores and distances between layers in the model hierarchy (Fig. 2B). 228

Inverse Linear Softplus increases specificity (Fig. 4B). The 229 230 reason is that, by improving predictivity for the same-layer, In-231 verse Linear Softplus improves a key component of specificity, identification of same-layer similarity across subjects, while 232 ²³³ maintaining inter-layer separation (Fig. 4C).

Although classical RSA (Representational Similarity Anal- 287 on the mouse data over linear regression. 234 ysis) is not a predictive mapping method (instead comparing 288 235 236 237 238 239 240 specificity (Fig. 4B). 241

Both very strict and very flexible methods impair speci-242 ficity. Soft matching, a strict method that matches individ-243 ual units between populations, is not only worse for predictiv-244 ity (Fig. 4A), but also for specificity (Fig. 4B) as evaluated 245 using the silhouette score. The reason is that soft match-246 ing (because of its low same-layer predictivity) has low same-247 layer identifiability and therefore low specificity (Fig. 4C). Soft 248 matching also has lower specificity as evaluated using hierar-249 chy correlation (Fig. 4B). This is because soft matching rates 250 adjacent layers, such as layers 2 and 3, as dissimilar even 251 compared to more distant layers, such as layers 2 and 4 (Fig. 252 4C). 253

At the other extreme, the flexible 7-layer MLP does not max-254 imize specificity, because it reduces inter-layer separation, 255 though the gain in same-layer identifiability slightly improves 256 its specificity over ridge regression. This result illustrates the 310 Bidirectional mapping can improve separation between 257 258 259 the MLP achieve similar levels of same-layer predictivity, but 313 assessed similarity to the mouse brain responses. 260 261 specificity. 262

263 264 265 266 (Fig. 4D). 267

Testing IATC candidates for a mouse population 268

269 270 averaged over 50 trials while the mice passively viewed 118 271 different visual stimuli. We evaluate methods for predictivity 272 273 274 models of the mouse brain. 275

Rank order of transform classes is largely similar be-276 tween mouse and model populations. The rank order of 277 ransform classes in terms of same-area predictivity provides 278 evidence for which transform classes are better IATC can-279 didates (despite the absolute scores being limited by rela-280 tively few response patterns for fitting the transforms, as well 281 as a limited neuronal sample). As in the model population, 282 soft matching achieves the lowest same-area predictivity, with 283 linear regression performing substantially better (Figure 5A). 284 Moreover, our biologically motivated transform classes that 285 account for the activation function further improve predictivity

The fact that transform classes that account for the activasummary statistics of population responses), we can still eval- 289 tion function are best for same-area predictivity hints at the uate it for specificity as a useful benchmark against which our 290 possibility that, just as in the simulated population, the pre-IATC candidates can be compared, as it is widely used for 291 non-linearity responses of two typical mouse subject are reneural response comparisons (Kriegeskorte et al., 2008). We 292 lated by a linear transform, and the relationship between postfind that Inverse Linear Softplus outperforms RSA in terms of 293 non-linearity responses is modified by the non-linearity. The 294 fact that Linear Softplus (which models the activation function as a softplus) performs about the same as Linear Non-295 296 linear (which models the activation function as an exponen-297 tial) means that we cannot tell, based on our results, which ²⁹⁸ of these activation functions better models the activation func-299 tions generating the mouse responses.

> 300 Strict methods that attempt to match individual units, such 301 as soft matching, are worse for specificity compared to ridge 302 regression and our biologically motivated transform classes 303 (Figure 5B). This confirms that low predictivity can lead to low 304 specificity by limiting same-area identifiability between sub-305 jects. Furthermore, this confirms that the most promising can-306 didate IATCs - Linear Nonlinear and Linear Softplus - are con-307 strained enough to separate responses from different brain 308 areas, in addition to being flexible enough to align same-area 309 responses between mouse subjects.

importance of identifying the strictest set of transforms that 311 brain models. We also evaluate how well each method maps accurately across subjects. Inverse Linear Softplus and 312 separates different candidate models with respect to their We Inverse Linear Softplus is more constrained, leading to higher 314 map 5 layers from four candidate models to the mouse re-315 sponses: the ReLU-based AlexNet model of mouse visual These results illustrate the utility of the IATC, as attempting 316 cortex (Nayebi et al., 2022), our noisy softplus version of that to identify it yielded a transform class that achieves high pre- 317 model, a ResNet model trained on ImageNet categorization dictivity and high specificity. In fact, our best IATC candidate 318 with 64x64 resolution stimuli, and a VGG-16 model trained on approaches the Pareto frontier for predictivity and specificity 319 ImageNet categorization with 224x224 resolution stimuli (un-320 like the low resolution mouse visual system). We apply the 321 same noise correction procedure for predictivity scores used 322 in (Nayebi et al., 2022) to account for trial-to-trial variability We now evaluate our methods on a mouse population, us- 323 (App. H). Model separation for a given area is evaluated as ing Neuropixels recordings for 31 subjects in 6 brain areas 324 the absolute difference in assessed brain similarity between 325 models (averaged over model pairs and model layers).

Typically, when mapping models to brains, model reand specificity when comparing responses between mouse 327 sponses are mapped to brain responses, but the other direcsubjects and also examine their ability to separate candidate 328 tion of mapping is not considered. Guided by the IATC, we 329 map bidirectionally, just as we do when aligning two mouse 330 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 332 333 brain data to models. When the scores for both mapping di-334 rections are averaged, the methods are all roughly compara-335 ble in terms of model separability. These results indicate that 336 stricter methods like soft matching are not generally better for 337 model separation. Furthermore, mapping in both directions 338 can increase separation between models compared to unidi-³³⁹ rectional mapping from model to brain.

> Overall, the results in this section show that our IATC re-340 341 sults 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.

342 343 344 for the IATC (such as Linear Nonlinear) are relatively good for 364 fMRI data obscures the effect of the non-linearity. 345 both prediction and specificity. Our results also highlight the 346 importance of the IATC's bidirectionality for model separation. 347

Testing IATC candidates on a human population 348

We evaluate transform classes on a large scale human fMRI 349 dataset, the Natural Scenes Dataset (Allen et al., 2021), and 350 evaluate methods for predictivity and specificity when map-351 ping between human subjects for 7 visual areas: V1, V2, V3, 352 hV4, as well as a higher area in each of the lateral, ventral, 353 and parietal streams. Finally, we use IATC-guided bidirec-354 tional mapping to better separate between models of the hu-355 man visual system. 356

Ridge regression achieves the best intra-area cross-sub- 377 357 358 359 360

real brain data. In particular, we find evidence that the IATC 361 rately across subjects (Figure 6A). Although Linear Nonlinear for the mouse population is shaped by the non-linear activa- 362 is close to ridge regression in terms of predictivity, it does not tion function. Moreover, we again find that viable candidates 363 do noticeably better, perhaps because the low resolution of

> Ridge regression improves specificity and visual hierar-365 chy identification. Under the mean-area silhouette score, 366 367 soft matching performs worse than other methods while ridge 368 regression performs best, suggesting that soft matching has lower specificity than ridge regression (Figure 6B). Ridge re-370 gression scores are more correlated with distances in the vi-371 sual hierarchy, suggesting that ridge regression better tracks 372 differences across the functional hierarchy. The improved hierarchical correlation for ridge regression compared to soft 373 374 matching is apparent on an MDS plot visualizing distances be-375 tween response profiles for different subjects and brain areas 376 (Figure 6C).

We also consider sparse regressions (Prince et al., 2024), ject predictivity. We again find that soft matching is unable 378 which use a lasso penalty to encourage sparse weights (with to map across subjects with high predictivity. A more flexi- 379 or without a positive weights constraint). These methods ble transform, ridge regression, is needed to map more accu- 380 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).

381 382 383 384 а 385 386 best IATC candidate. 387

Bidirectional IATC-guided mapping improves separation 388 of candidate brain models. Recent work (Margalit et al., 389 2023; Finzi et al., 2022) introduced a topographic model 390 (TDANNs) of the visual system that combines functional and 391 spatial constraints. While the TDANN with an intermediate 392 level of spatial loss strength $\alpha = 0.25$ predicted topograph-393 394 ical properties of visual cortex better than alternative models, a key question has been whether the TDANN's response 395 396 patterns quantitatively match neuronal responses better than

matching. We observe a loss in predictivity and specificity for 397 non-topographic models. Here, the issue of a correct comparsparse regressions (Fig. 6A,B), highlighting that even some- 398 ison method has been crucial, as unidirectional linear regreswhat stricter-than-linear methods can impair predictivity and 399 sion (from model to brain) failed to differentiate the TDANN specificity. Although we cannot compare ridge regression to 400 from non-topographic models, while a 1-1 mapping did (Marvery flexible control such as an MLP given the dataset size 401 galit et al., 2023, Fig. 6A,B). However, the very strictness of 1-(1000 response patterns), ridge regression seems to be the 402 1 mapping limited brain predictivity and the inter-animal noise 403 ceiling, leaving it unclear how strong the evidence is in favor of 404 the TDANN model. We therefore investigated whether IATCguided methods could distinguish between the TDANN and 405 406 alternative models in terms of matching neuronal responses.

> Guided by the IATC, we did ridge regression in both direc-407 408 tions between models and the brain. Although linearly map-409 ping model responses to the brain data does not separate ⁴¹⁰ strongly between the models, linearly mapping the brain data to the model responses separates strongly between the mod-₄₁₂ els (Figure 7A), identifying the TDANN with $\alpha = 0.25$ or 0.5 413 as being the most brain-like, convergent with soft matching re

414 sults and also with prior evidence in favor of that model (Mar- 468 ing to low identifiability and thus low specificity (Fig. 2A). On galit et al., 2023). In fact, model separation using bidirectional 469 each population, the best working estimate of the IATC im-415 ridge regression is much larger than for soft matching. This 470 proves same-area identifiability by mapping responses accu-416 result can be attributed to the fact that soft matching is an ex- 471 rately across subjects, while still maintaining inter-area sepa-417 tremely strict method, which results in low predictivity scores 472 ration, leading to high specificity. Thus, there is, in fact, no real 418 for all models (Figure 7A, right-most plot). By distinguish- 473 tension between specificity and predictivity. 419 ing more strongly between TDANN ($\alpha = 0.25.5$) and alterna- 474 420 421 422 stronger evidence in favor of the TDANN. 423

424 425 426 427 428 429 guided bidirectional mappings. 430

Discussion

431

432 433 needed to map neural responses accurately between animal 489 pling of neurons may affect the brain-to-model direction. 434 subjects in a population (for the same brain area). We find in 490 435 436 437 438 for a model-brain comparison method. In a simulated popu- 494 439 lation of neural networks, we identified how the neuronal acti- 495 accurately describes inter-subject variability. 440 vation function shapes the IATC. leading to a transform class 496 441 442 mapping classes. On a mouse electrophysiology dataset, we 498 443 also find evidence that the IATC is constrained by the neu- 499 the IATC in a data-driven fashion using large-scale optimiza-444 ronal activation function, suggesting that IATC results for the 500 tion techniques. An especially exciting possibility will be to 445 model population can meaningfully generalize to real brain 501 use such data-driven estimation methods to strengthen and 446 data. On a human dataset, the resolution of the fMRI data 502 refine the preliminary results we show here suggesting that 447 does not allow us to observe an effect of the activation func- 503 neural circuit mechanisms constrain the IATC. This will be an 448 tion on the IATC, but we still see differentiation between can- 504 increasingly realistic prospect as neural datasets grow in size 449 didate transform classes that is consistent with our findings 505 (with many subjects, neurons per subject, and stimuli). 450 from simulated population and mouse data. Moreover, we 506 451 use the IATC-guided bidirectional mappings to enable better 507 the brain, have privileged axes in their representations: unit-452 453 454 non-topographic models. 455

456 specificity and predictivity would likely be in tension - with 512 class, similar to the Linear Nonlinear transform class that mod-457 458 459 460 461 462 463 464 465 arating responses of different types. Extremely strict methods 521 data-driven discovery and optimization methods. 466 fail to align same-area responses well across subjects, lead-467

A key aspect of the IATC approach is to map bidirectionive models such as non-topographically constrained models 475 ally between models and brains, not just in the model to brain $(\alpha = 0)$, IATC-guided bidirectional ridge regression provides 476 direction, just as when comparing two brains to each other. 477 Considering both directions can reveal cases where a given Along similar lines, Khosla & Williams (2023) observed 478 model contains spurious features, improving model separacases where linearly mapping models to brain responses did 479 tion. Unlike previous works that motivated symmetry with the not separate models, but soft matching did. Revisiting this 480 assumption that similarity scores must be distance metrics analysis but with bidirectional mappings, we found that bidi- 481 (Williams et al., 2021; Khosla & Williams, 2023), under our rectional ridge regression increased model separation over 482 IATC approach, similarity naturally emerges from bidirectional soft matching (Fig. 7B), further confirming the utility of IATC- 483 relationships between brains in a population. Our approach 484 also differs from work that treated both directions of mapping as separate, potentially inconsistent methods (Soni et al., 485 486 2024) rather than as two components of a single method. In The IATC provides a principled framework for model-brain 487 future work, we plan to further investigate bidirectional mapcomparison by identifying the strictest set of transforms 488 pings and address potential limitations, such as how subsam-

A limitation of our results using a simulated population is three settings (model population, mouse population, and hu- 491 the question of whether the sources of variability we consider man population) that a working estimate of the IATC achieves 492 (different seeds for weight initialization and training data order) both high predictivity and high specificity, two key desiderata 493 are anything like sources of actual brain variation. In future work, we will work towards a "generative model" that more

A second direction for improvement will be to improve IATC hat improves predictivity and specificity relative to standard 497 estimation. Rather than evaluating a small set of transform classes as done in this paper, we hope to systematically learn

Recent work has shown that neural networks, and likely model-brain comparisons, uncovering new evidence differen- 508 level tuning curves that are similar between subjects, and tiating topographic models of the visual system compared to 509 which cannot be linearly remixed without constraint (Khosla & 510 Williams, 2023). A natural synthesis of this result with our find-When beginning this work, we assumed that the goals of 511 ings would be a hybrid "Linear Subspace-Nonlinear" transform stricter methods (such as soft matching) being better for speci- 513 els the activation function, but where the linear portion of the ficity of model identification and worse at prediction, and more 514 transform is constrained by the preferred axes to specific subflexible methods (such as linear regression) showing the op- 515 spaces of allowable transforms. This hybrid would be stricter posite pattern (Ding et al., 2021; Kornblith et al., 2019; Con- 516 than full Linear-Nonlinear, but more flexible than soft matchwell et al., 2022; Finzi et al., 2022). However, the intuition 517 ing. We hypothesize that such a transform class would be a that stricter methods are generally better for specificity over- 518 better estimate of the true IATC and would occupy a "sweet looks the fact that specificity requires identifying high similarity 519 spot" on the strictness-flexibility continuum (Fig. 8). Actually across subjects for responses of the same type, not just sep- 520 estimating these transform subspaces will likely involve novel



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.

522

Acknowledgements

This work was supported by the following awards: To I.T.: 523 Patrick Suppes Philosophy of Science Dissertation Award. To 524 D.L.K.Y.: Simons Foundation grant 543061, National Science 525 Foundation CAREER grant 1844724, National Science Foun-526 dation Grant NCS-FR 2123963, Office of Naval Research 527 grant S5122, ONR MURI 00010802, ONR MURI S5847, and 528 ONR MURI 1141386 - 493027. We also thank the Stanford 529 HAI, Stanford Data Sciences and the Marlowe team, and the 558 530 531

Google TPU Research Cloud team for computing support.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., 533 Citro, C., ... Zheng, X. (2015). TensorFlow: Large-scale 534 535 machine learning on heterogeneous systems. Retrieved from https://www.tensorflow.org/ (Software avail-536 able from tensorflow.org) 537

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., 538 Dowdle, L. T., ... Kay, K. (2021, December). A massive 7t 539 fmri dataset to bridge cognitive neuroscience and artificial 540 intelligence. Nature Neuroscience, 25(1), 116-126. Re-541 trieved from http://dx.doi.org/10.1038/s41593-021 542

-00962-x doi: 10.1038/s41593-021-00962-x 543

532

Canatar, A., Feather, J., Wakhloo, A., & Chung, S. (2024). 544 A spectral theory of neural prediction and alignment. Ad-545 vances in Neural Information Processing Systems, 36. 546

- Cao, R., & Yamins, D. (2021). Explanatory models in neu-547 roscience: Part 1-taking mechanistic abstraction seriously. 548 arXiv preprint arXiv:2104.01490. 549
- Chichilnisky, E. (2001). A simple white noise analysis of neu-550 ronal lightresponses. Network: computation in neural sys-551 tems, 12(2), 199. 552
- Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, 553 T. (2022). What can 1.8 billion regressions tell us about the 554 pressures shaping high-level visual representation in brains and machines? BioRxiv, 2022-03. 556

Ding, F., Denain, J.-S., & Steinhardt, J. (2021). Grounding rep-557 resentation similarity through statistical testing. Advances in Neural Information Processing Systems, 34, 1556–1568. 559

Feather, J., Leclerc, G., Mkadry, A., & McDermott, J. H. 560 (2023). Model metamers reveal divergent invariances be-561 tween biological and artificial neural networks. Nature Neu-562

- roscience, 26(11), 2017-2034. 563
- Finzi, D., Margalit, E., Kay, K., Yamins, D. L., & Grill-Spector, 564 K. (2022). Topographic dcnns trained on a single self-565
- supervised task capture the functional organization of cor-566
- tex into visual processing streams. In Svrhm 2022 work-567 shop@ neurips. 568

Gao, X., Wen, M., Sun, M., & Rossion, B. (2022). A genuine 569 interindividual variability in number and anatomical localiza-570

tion of face-selective regions in the human brain. Cerebral 571 Cortex, 32(21), 4834-4856. 572

Haug, H. (1987). Brain sizes, surfaces, and neuronal sizes 573

- of the cortex cerebri: a stereological investigation of man 574
- and his variability and a comparison with some mammals 575
- (primates, whales, marsupials, insectivores, and one ele-576
- phant). American Journal of Anatomy, 180(2), 126-142. 577

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, 578

- S. V., & McDermott, J. H. (2018). A task-optimized neu-579
- ral network replicates human auditory behavior, predicts 580
- brain responses, and reveals a cortical processing hierar-581
- chy. Neuron, 98(3), 630-644. 582

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep su-583 pervised, but not unsupervised, models may explain it cor-584

- tical representation. PLoS computational biology, 10(11), 638 585 e1003915. 639 586
- Khosla, M., & Williams, A. H. (2023). Soft matching distance: 640 587
- A metric on neural representations that captures single- 641 588 neuron tuning. arXiv preprint arXiv:2311.09466. 589 642
- Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Sim- 643 590
- ilarity of neural network representations revisited. In Inter- 644 591

592

- Kriegeskorte, N., & Diedrichsen, J. (2016). Inferring brain- 646 593
- computational mechanisms with models of activity mea-647 594
- surements. Philosophical Transactions of the Royal Society 648 595 B: Biological Sciences, 371(1705), 20160278. 649 596
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Rep- 650 Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, 597 resentational similarity analysis-connecting the branches of 651 598
- systems neuroscience. Frontiers in systems neuroscience, 652 599 4. 653 600
- Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., 654 601
- & Yamins, D. L. (2023). A unifying principle for the functional 655 602 organization of visual cortex. bioRxiv. 603
- McCullagh, P., & Nelder, J. (2019). Generalized linear models. 604 Routledge. 605
- 659 Mineault, P., Bakhtiari, S., Richards, B., & Pack, C. (2021). 606
- Your head is there to move you around: Goal-driven models 607 of the primate dorsal pathway. Advances in Neural Informa- 661 608
- tion Processing Systems, 34, 28757-28771. 609
- 663 Navebi, A., Attinger, A., Campbell, M., Hardcastle, K., Low, I., 610
- Mallory, C. S., ... others (2021). Explaining heterogeneity 664 611 in medial entorhinal cortex with task-driven neural networks. 665
- 612 Advances in Neural Information Processing Systems, 34, 666 613
- 12167-12179. 614 667
- 668 Nayebi, A., Kong, N. C., Zhuang, C., Gardner, J. L., Norcia, 615
- A. M., & Yamins, D. L. (2022). Mouse visual cortex as 669 616
- a limited resource system that self-learns an ecologically- 670 617 general representation. bioRxiv, 2021-06. 671 618
- Nayebi, A., Rajalingham, R., Jazayeri, M., & Yang, G. R. 672 619
- (2024). Neural foundations of mental simulation: Future 673 620
- prediction of latent representations on dynamic scenes. Ad- 674 621 675
- vances in Neural Information Processing Systems, 36. 622
- Olshausen, B. A., & Field, D. J. (1996). Emergence of simple- 676 623
- cell receptive field properties by learning a sparse code for 677 624 natural images. Nature, 381(6583), 607-609. 678 625
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., 679 626
- 627 learn: Machine learning in Python. Journal of Machine 681 628
- Learning Research, 12, 2825-2830. 682 629 683
- Prince, J. S., Conwell, C., Alvarez, G. A., & Konkle, T. (2024). 630
- A case for sparse positive alignment of neural systems. In 631
- Iclr 2024 workshop on representational alignment. 632
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the 633
- interpretation and validation of cluster analysis. Journal of 634
- computational and applied mathematics, 20, 53-65. 635
- Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, 636
- E. A., Kanwisher, N., ... Fedorenko, E. (2021). The 637

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.
- national conference on machine learning (pp. 3519–3529). 645 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.
 - 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., 657 Schmidt, M.-A., Bittarello, L., ... pwojtasz (2025).Retrieved from https://github.com/ Quantco/glum. Quantco/glum

658

662

- 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.
- Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit- 680 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.

684 Appendix

685 A Central hypercolumn selection

⁶⁸⁶ In order to map between units with similar functional roles (at least for the same model layer), we do our model-model fits using ⁶⁸⁷ 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 ⁶⁸⁸ image). Indeed, even when constraining the mapping to use only the central hyper-column, we are able to identify high similarity ⁶⁸⁹ across model instances for the same layer, at least when assessing pre-non-linearity responses using a linear transform.

690 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:

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

⁶⁹¹ where **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 ⁶⁹² $1/N_X$, and **C** is the matrix of Pearson correlations between each source neuron and each target neuron. The transport matrix ⁶⁹³ can be interpreted as a joint probability distribution over source neurons and target neurons (where the marginal distributions ⁶⁹⁴ are uniform discrete). Thus, the soft matching score is the *expected* correlation between source and target neurons, according ⁶⁹⁵ 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:

$$\mathbf{\hat{Y}}_{j} = rac{\mathbf{\sigma}(\mathbf{Y}_{j})}{\mathbf{\sigma}(\mathbf{X}_{i})} [\mathbf{X}_{i} - \mathbf{\bar{X}}_{i}] \mathbf{C}_{ij} + \mathbf{\bar{Y}}_{j}$$

⁶⁹⁶ This is essentially using the correlation 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:

$$\mathbf{\hat{Y}}_{j} = N_{Y} \mathbf{\sigma}(\mathbf{Y}_{j}) \Sigma_{i} \frac{\mathbf{X}_{i} - \mathbf{X}_{i}}{\mathbf{\sigma}(\mathbf{X}_{i})} \mathbf{T}_{ij} \mathbf{C}_{ij} + \mathbf{\bar{Y}}_{j}$$

⁶⁹⁷ C Motivating the softplus activation function with a simple model of a noisy spiking process

⁶⁹⁸ Our simulation of spike counts is based on the following highly simplified model. We assume that a neuron receives total input ⁶⁹⁹ $X \sim \mathcal{N}(\mu, \sigma^2)$ and that during a single time interval equal to the neuron's refractory period (which we assume to be about 1 ms), ⁷⁰⁰ 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 ⁷⁰¹ 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:

$$softplus(x) = ln(1 + e^x)$$

702 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.

⁷¹⁰ To avoid numerical difficulties for small values of $k = \lambda$, we scale the softplus outputs by 100 before sampling from the Gamma ⁷¹¹ distribution. We then train the noisy softplus models so that their instance recognition training score (as well as validation score ⁷¹² on ImageNet categorization) are equal to those of the ReLU-based AlexNet models.

713 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:

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

⁷¹⁴ In practice, to avoid numerical difficulties for very small values of *y*, we do not apply this formula directly and instead use a more ⁷¹⁵ numerically stable implementation of the softplus inverse adapted from the TensorFlow library (Abadi et al., 2015).

716 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 717 can directly invert the activation function to recover the pre-non-linearity responses. However, when mapping animals to animals 718 (or, if enough neurons are measured, animals to models), we cannot easily invert the activation function if we do not know 719 its exact form for a given neuron. Yeo-Johnson scaling uses a power transformation to make the features closer to normally 720 distributed over the stimuli. We expect this transformation to make the post-non-linearity features more correlated with pre-non-721 linearity responses because the non-linear activation function skews the distribution of the pre-non-linearity responses (which 722 are roughly normally distributed over the stimuli). Indeed, we find that Yeo-Johnson scaling noticeably increases the Pearson 723 correlation (Fig. 10) with the pre-non-linearity responses for the noisy softplus models, almost as much as if you had directly 724 applied the inverse of the softplus activation function to the post-non-linearity responses. We hypothesize that Yeo-Johnson 725 scaling has a similar effect in the case of animal firing rates. 726

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.

731 G Implementation details of GLMs

⁷³² The GLM object is created using the *glum* package (Thompson et al., 2025). Each GLM specifies the inverse link function that ⁷³³ 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.

⁷³⁴ in the response variable (Poisson noise in the case of Linear Nonlinear or Linear Softplus). The weights of the specified GLM
 ⁷³⁵ are then optimized through Iterative Reweighted Least Squares.

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

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

⁷³⁶ 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 ⁷³⁷ (trial averaged responses of the source model after applying either Yeo-Johnson scaling or exactly inverting the activation func-⁷³⁸ tion) and θ is the fitted linear weights. When predicting noisy softplus model responses, we set c = 100, the same softplus output ⁷³⁹ scaling we used when training the models themselves. But when fitting Linear Softplus to predict mouse responses, we do not ⁷⁴⁰ know *a priori* the optimal scaling parameter and must cross-validate values of *c* along with the ridge penalty using GridSearchCV ⁷⁴¹ in sklearn.

742 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-totrial 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:

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

⁷⁴³ where \mathcal{M} is a given mapping method (such as ridge regression), r_{train} is the model responses in a given layer (which, except ⁷⁴⁴ for the noisy softplus models, are deterministic) over the training stimuli, t_{train}^B is the *true* trial-averaged (averaged over the ideal ⁷⁴⁵ limit of infinitely many trials) responses of a particular subject and brain area *B* over training stimuli, $\mathcal{M}(r_{\text{train}}; t_{\text{train}}^B)_{\text{test}}$ are the ⁷⁴⁶ test predictions under the mapping method of the target animal's responses over the test stimuli using the model responses as ⁷⁴⁷ predictors, and t_{test}^B are the actual ground-truth responses of the target animal over test stimuli. This quantity cannot be directly ⁷⁴⁸ computed because we do not have infinitely many trials per stimulus, and instead must estimate it based on finitely many trials ⁷⁴⁹ (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:

$$\operatorname{median} \left\langle \frac{\operatorname{Corr} \left(\mathcal{M}(r_{\operatorname{train}}^{\ell}; s_{1, \operatorname{train}}^{B})_{\operatorname{test}}, s_{2, \operatorname{test}}^{B} \right)}{\sqrt{\widetilde{\operatorname{Corr}} \left(\mathcal{M}(r_{\operatorname{train}}^{\ell}; s_{1, \operatorname{train}}^{B})_{\operatorname{test}}, \mathcal{M}(r_{\operatorname{train}}^{\ell}; s_{2, \operatorname{train}}^{B})_{\operatorname{test}} \right) \times \widetilde{\operatorname{Corr}} \left(s_{1, \operatorname{test}}^{B}, s_{2, \operatorname{test}}^{B} \right)} \right\rangle} \right\rangle$$

⁷⁵⁰ where the median is computed over the target animal's neurons, and the $\langle ... \rangle$ represents an average over all bootstrap samples. ⁷⁵¹ The Corr represents a Spearman-Brown corrected Pearson correlation rather than a raw Pearson correlation. $s_{i, \text{ train/test}}^{B}$ repre-⁷⁵² sents 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 ⁷⁵³ *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.

757 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:

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

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.