Supporting Information: A connectivity-constrained computational account of topographic organization in high-level visual cortex

Nicholas M. Blauch[1,2] Marlene Behrmann[2,3] David C. Plaut[2,3]
[1]Program in Neural Computation [2]Neuroscience Institute [3]Department of Psychology
Carnegie Mellon University
{blauch,behrmann,plaut}@cmu.edu

November 9, 2021

1 Final model performance

We measured overall recognition performance of the main ITN model throughout training. The results are plotted for training and validation images as a function of training epoch in Figure S1. Additionally, we plotted the final accuracy for each domain as a function of processing time step in the model.

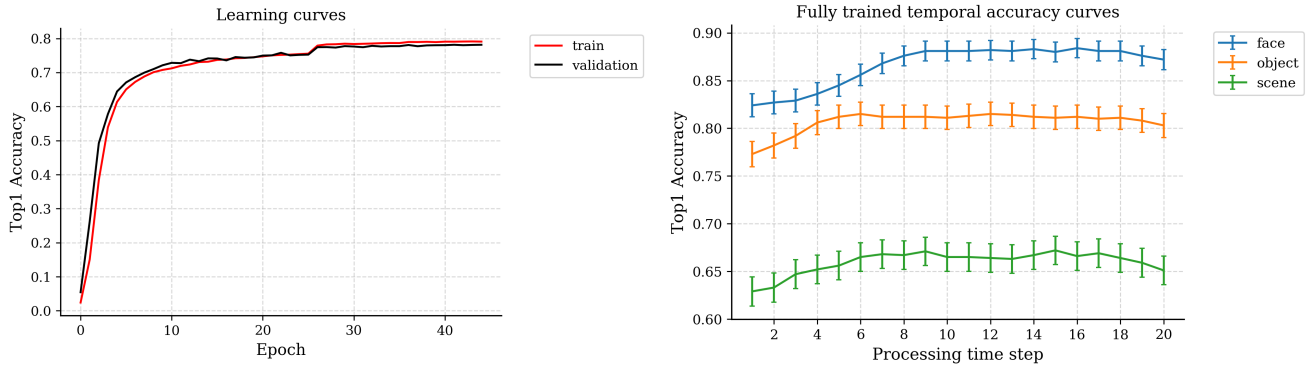


Figure S1: Performance of the main model. **A.** Overall accuracy throughout the training of IT using a pre-trained encoder. **B.** Final accuracy for each domain, plotted over model processing time steps.

2 A more complete assessment of domain-level functional organization

In the main paper, for clarity, we plotted lesion deficits at two specific lesion sizes. Here, for completeness, we plot both raw accuracy values and lesion deficits for a large range of lesions, shown in Figure S2.

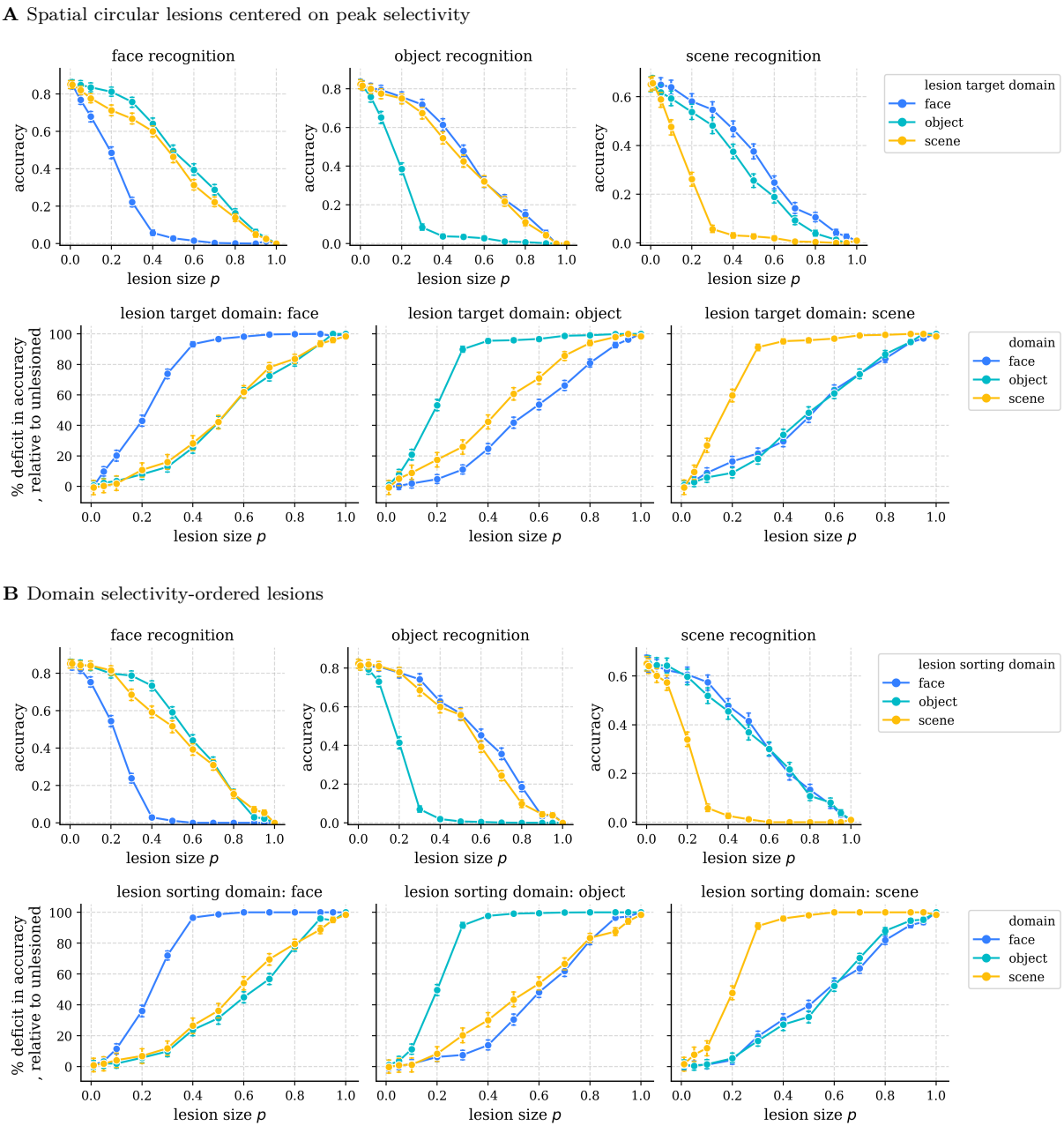


Figure S2: Lesion results in the ITN model. **A.** Damage from various sizes of circular topographic lesions centered on the peak of smoothed selectivity for each domain. **B.** Damage from various sizes of non-topographic lesions chosen by sorting units according to their selectivity for each domain. As the selectivity is highly topographic, the lesion masks in the topographic vs. selectivity-ordered lesion plots are very similar; however, for selective regions that are non-circular (such as that seen for objects), the selectivity-ordered lesion provides a more precise way of assaying the selective region.

To better understand the degree of representational competition and cooperation, we compare searchlight accuracy and mean readout maps across domains. The results are plotted in Figure S3 as scatter plots over units, colored by their selectivity between the two plotted domains, for all pairs of domain. The results corroborate our earlier findings,

demonstrating a strong but graded degree of specialization. In particular, we note the large degree of correlation between searchlight accuracies for objects and for scenes. This result is somewhat surprising given the relatively specialized effects of lesions, demonstrating that object and scene information may co-mingle but still be read out from largely different sets of units in order to optimize task performance.

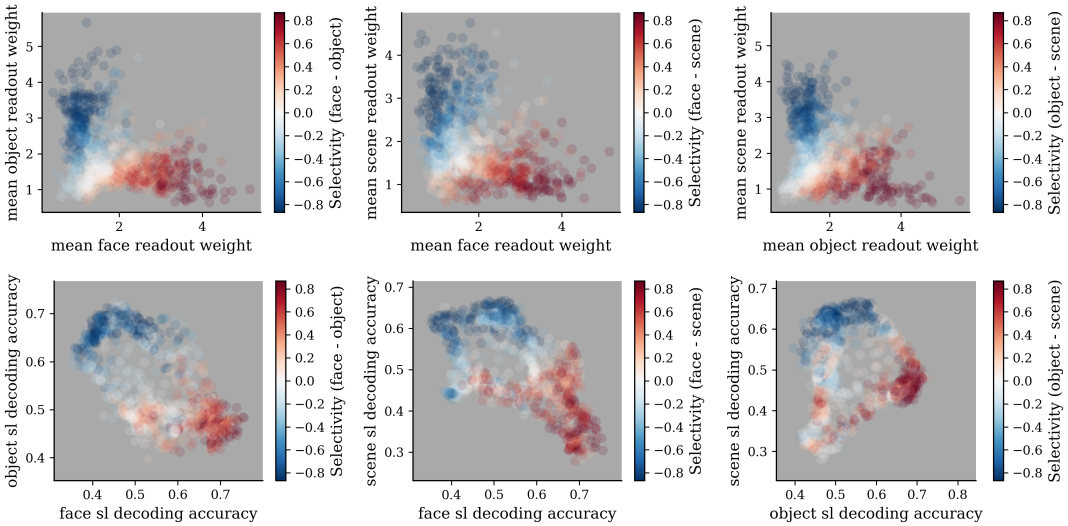


Figure S3: Graded competition and cooperation between domains in aIT. Top: readout weights into each domain, bottom: searchlight accuracy for decoding within each domain.

How consistent is the topographic responsiveness of categories within a domain? Within a given domain-selective area, are the weaker responses to non-preferred domains roughly uniform and weak across categories, or do some categories elicit notably stronger responses? To answer this question, we plotted the mean aIT response and readout weights of 10 randomly selected categories from each domain, overlaying the contour of significant smoothed domain-selectivity ($p<0.001$; smoothing performed as averaging selectivity over the 5% nearest units). These results are shown in Figure S4. Indeed, individual categories tend to elicit the strongest responses and largest readout weights within their respective domain-selective area, but vary in the specific within-area topography, as well as outer-area topography. In particular, we note that many of the object categories have significant but localized readout weight mass outside the object selective area.

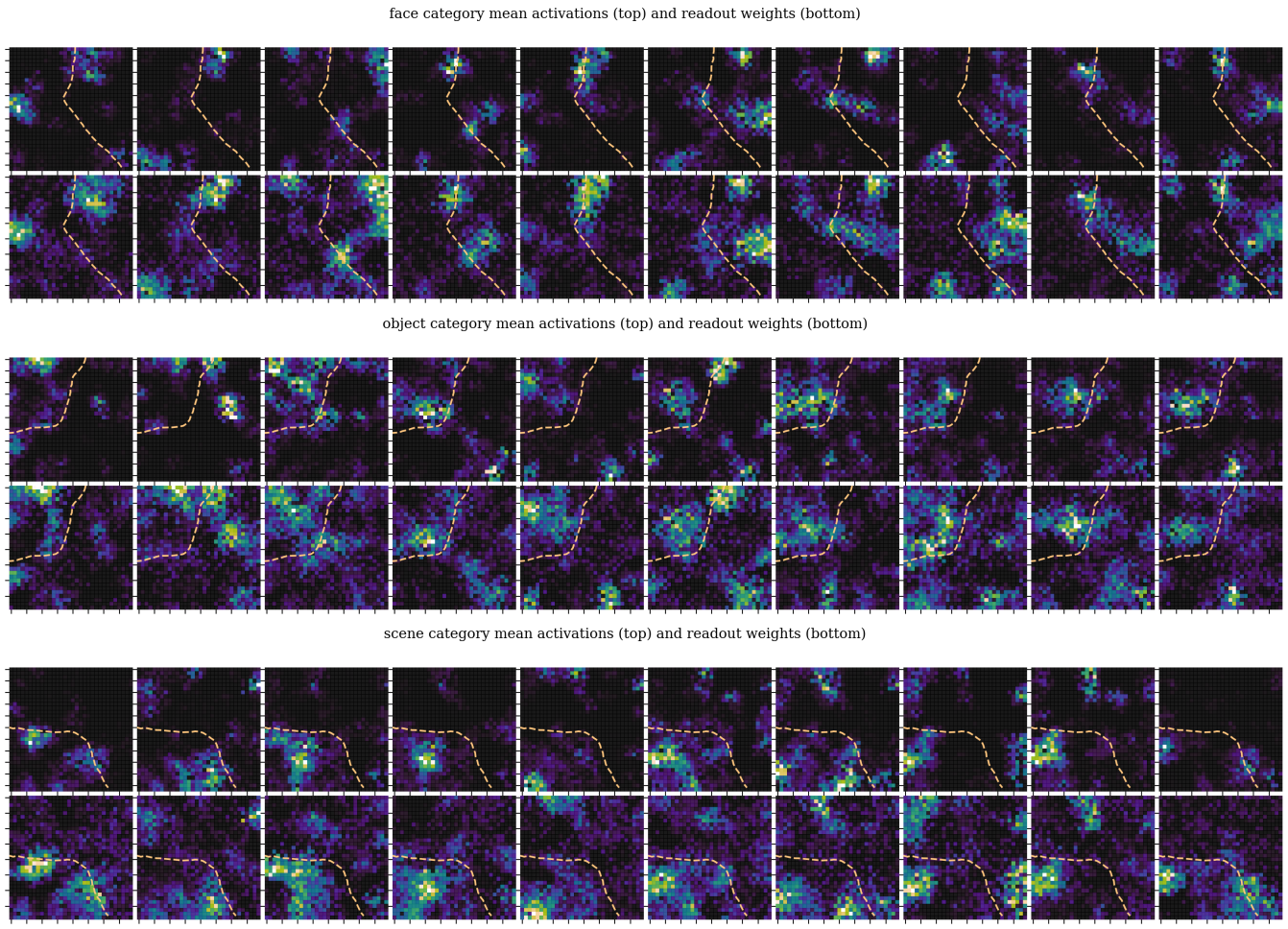


Figure S4: Category-level mean responses and readout weights for 10 example categories from each domain. For each domain, the top row shows mean category-level responses in aIT, and the bottom row shows the readout weights for the corresponding categories. Dashed yellow lines indicate the contour of statistically significant smoothed domain selectivity for the plotted category's domain ($p<0.001$; smoothing performed as averaging selectivity over the 5% nearest units).

3 Investigating responses to two common fMRI localizers

For completeness, in addition to the validation images used to compute domain selectivity in the main text, here we use two common localizers from the Konkle and Grill-Spector labs to compute category-selective maps. Figure S5A. replots the selectivity for the validation images from face, object, and scene domains; Figure S5B and C plot selectivity to each category in the Konkle and Grill-Spector localizers, respectively. Face selective topography is highly consistent across localizers, whereas object and scene selective topographies show somewhat greater variability, but still generally conforming to the large-scale organization shown in Figure S5A. The greater consistency of selectivity to faces may be because faces are more homogenous visually than objects and scenes. Other categories, such as bodies, characters, and scrambled images elicited patchy topographic responses.

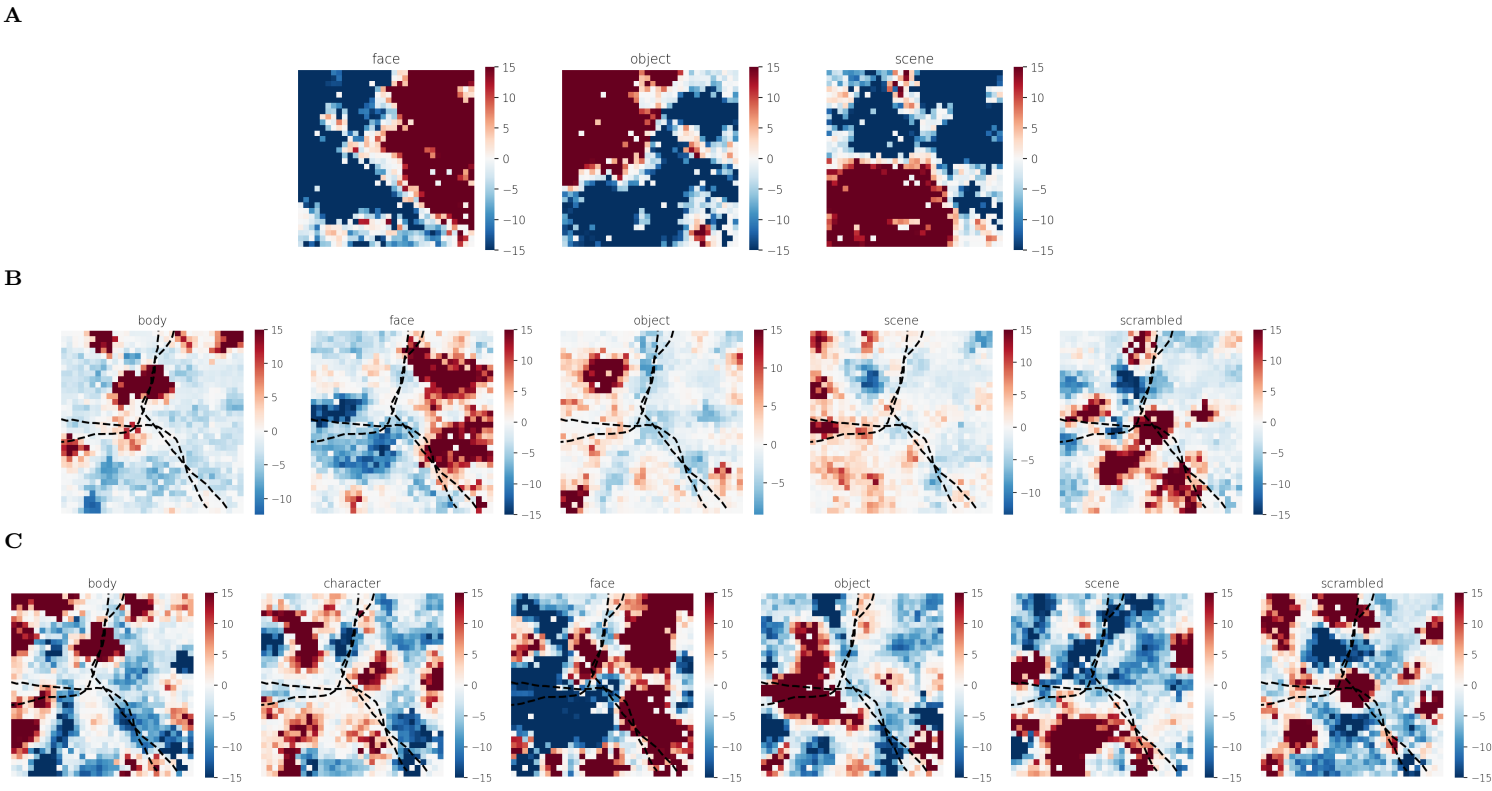


Figure S5: Selectivity to multiple stimulus sets. **A.** validation images used throughout the main text, for reference. **B.** A localizer set from the Konkle lab. **C.** The *fLoc* stimulus set from the Grill-Spector lab (Stigliani et al., 2015). Contours in **B** and **C** show outlines of significant smoothed domain selectivity ($p<0.001$; smoothing done by averaging using 5% nearest units).

4 Complete tuning analysis for the main model

Here we plot a finer-grained view of the tuning analysis for the main model (E/I EFF RNN), showing domain-selective and generic topography as well as the metrics computed per layer and celltype, shown in Figure S6.

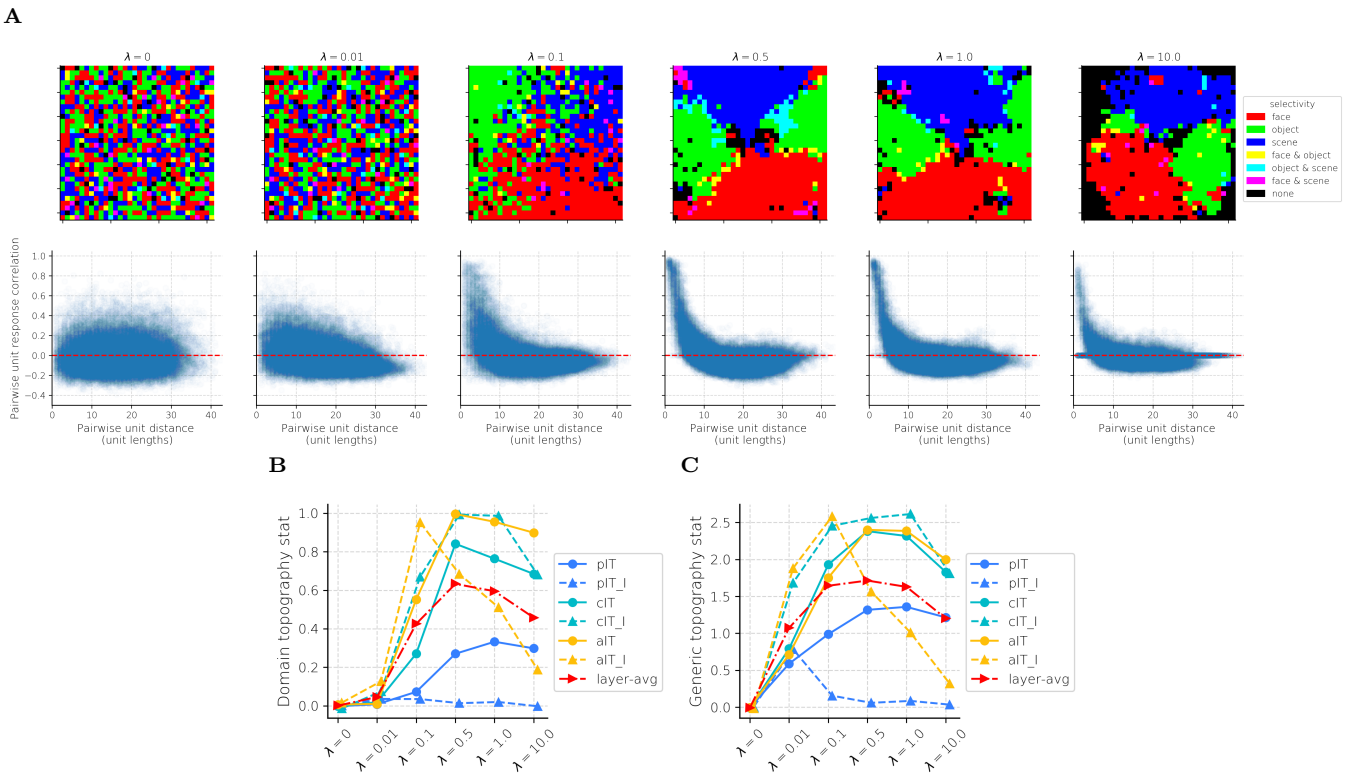


Figure S6: Domain-level and generic topography as a function of λ_w in the main model. A random seed of 1 was used for this analysis and the optimal λ_w was selected in order to maximize the layer average of the generic topographic organization statistic T_g. The main model used in the paper then used a random seed of 2 at this value of λ_w. **A.** Emergent topography. **D.** Quantification of generic topography (T_g, see Methods). **C.** Quantification of domain-level topography(T_d.)

5 Sparsification induces minimal performance deficits

In the main text we used an unweighted wiring cost $\mathcal{L}_{w,u}$ that first sparsified the network to eliminate the 99% weakest connections. In Figure S7 we demonstrate that in networks with a strong spatial penalty λ_w, this sparsification induces minimal performance deficits compared to the full network.

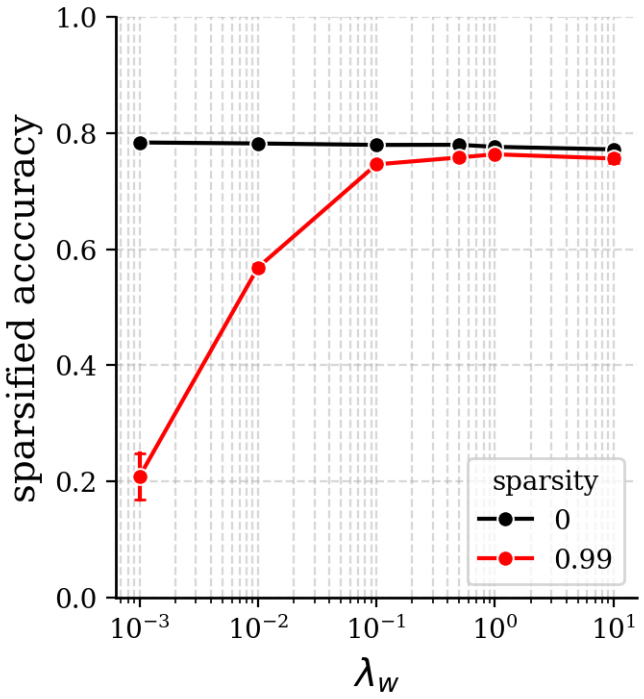


Figure S7: Comparison of final accuracy on the validation set for the main model and a sparsified version of it used to compute the wiring cost $\mathcal{L}_{w,u}$.

6 Topographic clustering of categories in aIT

In the main text, we used principal components analysis (PCA) to analyze the topographic organization of an ITN model. An alternative to this approach is to look for clusters of activation, which may be more localized than the global dimensions found by standard PCA. Here, we additionally perform a k-means clustering analysis of the readout weights in aIT, which can be easily visualized. Overall, the results corroborate the PCA results presented in the main text to indicate that topographic organization extends beyond the domain-level to encompass interpretable within-domain organization.

6.1 Clustering of object categories

Readout weight clustering for object categories is plotted in Figure S8. Each cluster is topographically localized and distinct from the other clusters. Moreover, each cluster is interpretable. The 0-th cluster contains nearly all of the inanimate categories, the 1st cluster contains animate non-mammalian categories, and the 2nd cluster contains mammals (including nearly all of the dogs).

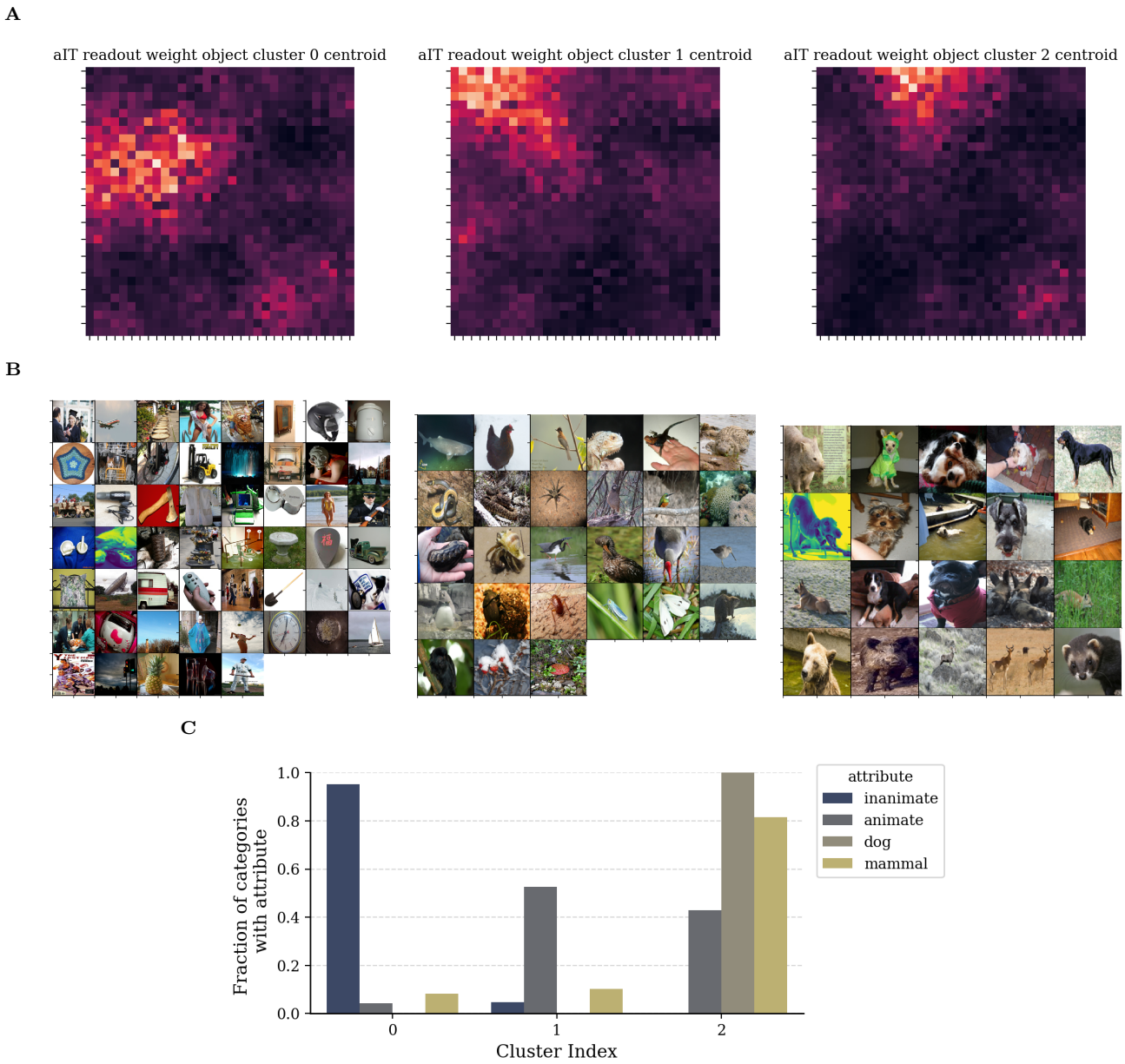


Figure S8: Clustering object category readout weights from aIT. $k = 3$ clusters were used in a K-means++ clustering algorithm that clustered the vectors of aIT readout weights over all 100 scene categories. **A.** The centroids of the clusters were then reshaped into the 2D aIT coordinates and visualized as heat maps. **B.** Cluster category members. **C.** Object attribute quantification.

6.2 Clustering of face categories

Readout weight clustering for face categories is shown in Figure S9. For faces, we do not plot category member images in order to preserve privacy. As for objects, topographically organized clusters of categories are found. In order to interpret these clusters, we labeled the attributes of biological sex and whether the modal hair color of an individual is blonde (inferred from a sample of 8 images per identity).The 0-th cluster contained mostly blonde females, the 1st cluster contained mostly non-blonde males, and the 2nd cluster contained mostly non-blonde females. As with the other domains, further analysis of other possible attributes (e.g. ethnicity, age) is possible, but is beyond the scope of this work.

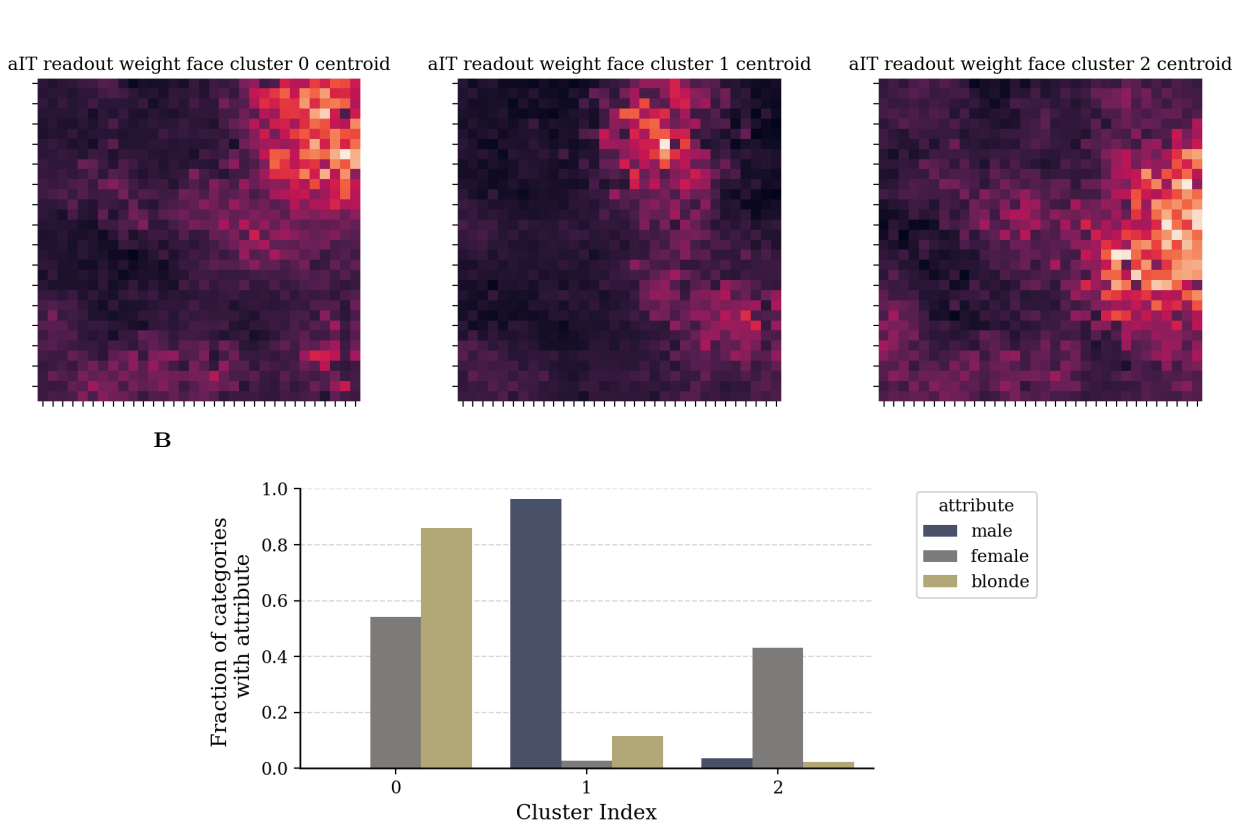


Figure S9: Clustering face category readout weights from aIT. $k = 3$ clusters were used in a K-means++ clustering algorithm that clustered the vectors of aIT readout weights over all 100 face categories. **A.** The centroids of the clusters were then reshaped into the 2D aIT coordinates and visualized as heat maps. **B.** Quantification of biological sex over the members of each cluster, revealing quantitative characteristics of each cluster.

6.3 Clustering of scene categories

Last, we performed clustering of scene categories. Again, topographically organized cluster centroids were discovered. To interpret these clusters, we labeled whether scenes were indoor or outdoor, and for outdoor, whether they were natural or manmade. We found that the 0th cluster contained outdoor scenes that were mostly natural, the 1st cluster contained indoor scenes, and the 2nd cluster contained mostly man-made outdoor scenes.

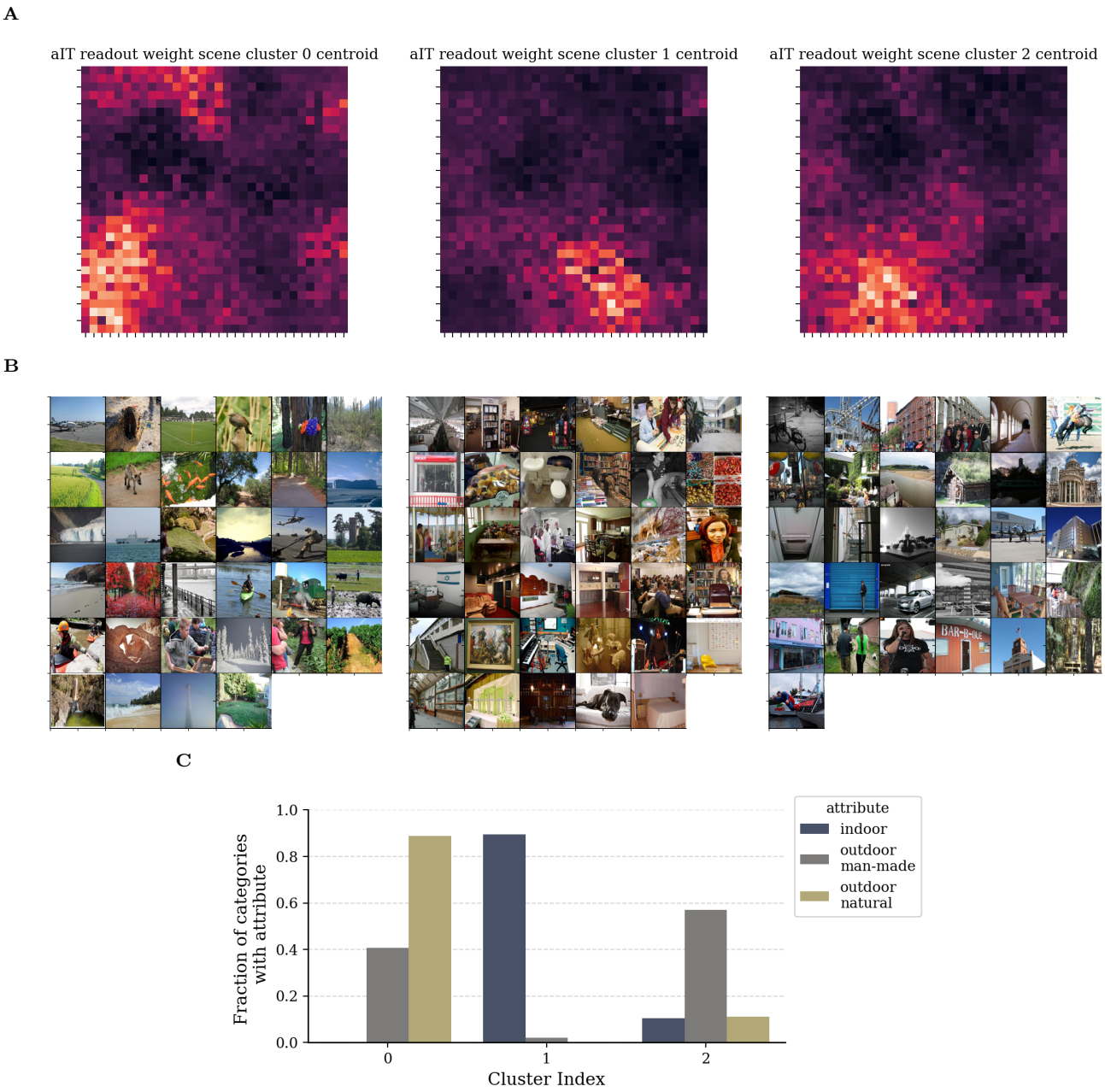


Figure S10: Clustering scene category readout weights from aIT. $k = 3$ clusters were used in a K-means++ clustering algorithm that clustered the vectors of aIT readout weights over all 100 scene categories. **A.** The centroids of the clusters were then reshaped into the 2D aIT coordinates and visualized as heat maps. **B.** Cluster category members. **C.** Scene attribute quantification.

7 E/I columns in aIT

In the paper we focused on E/I columns in cIT, where I cell responses were stronger than in aIT. The weakening of inhibition in aIT appears to occur in models with stronger λ_w, as the network discovers that it can reduce inhibitory weights (and thereby spatial costs) in the final layer in particular, where units project onto readout units subject to a squashing softmax nonlinearity. Here, for completeness, we plot the E/I columnar relationship in aIT.

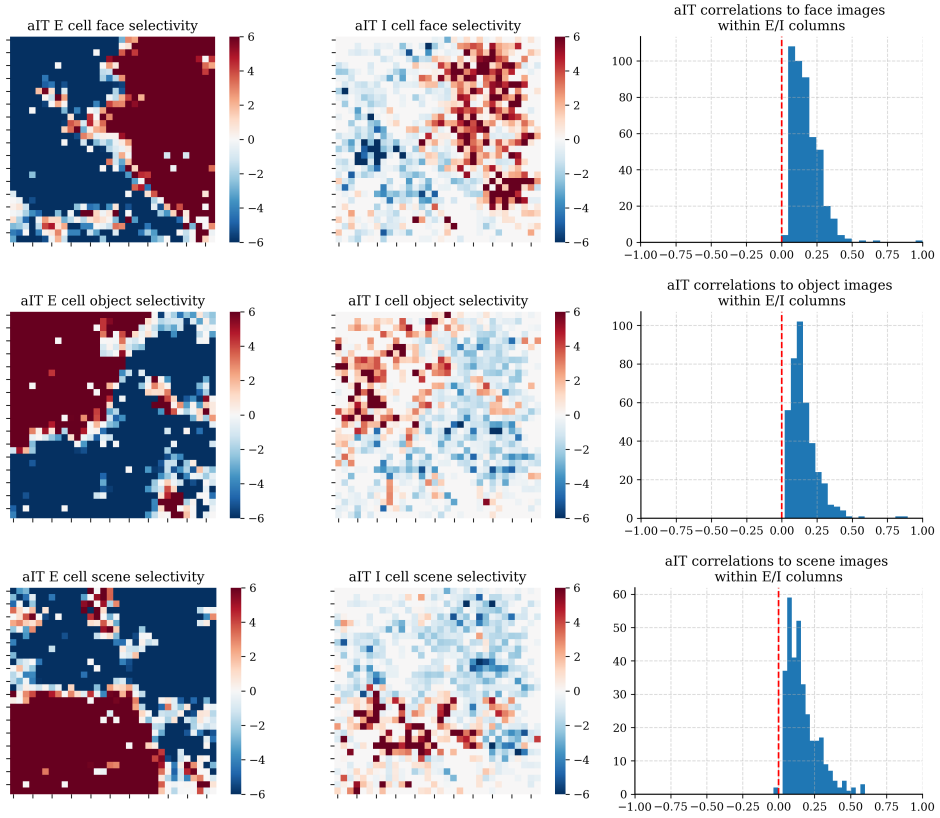


Figure S11: E/I columnar response selectivity in cIT. The selectivity for faces, objects, and scenes, respectively, is plotted for both E and I cells, and the histogram of correlations between E and I cells over all images of the given domain is computed over all locations on the grid, revealing dominantly positive correlations.

8 E/I columns depend on the spatial constraint

Do E/I columns occur spontaneously or due specifically to the constraint of minimal wiring costs? To answer this question, we plot E and I cell selectivities and E/I columnar correlations in a model with $\lambda_w = 0$ in Figure S12. Columns do not appear in this model, implying that the spatial cost term is a direct cause in the development of local coupling between columns of E and I cells.

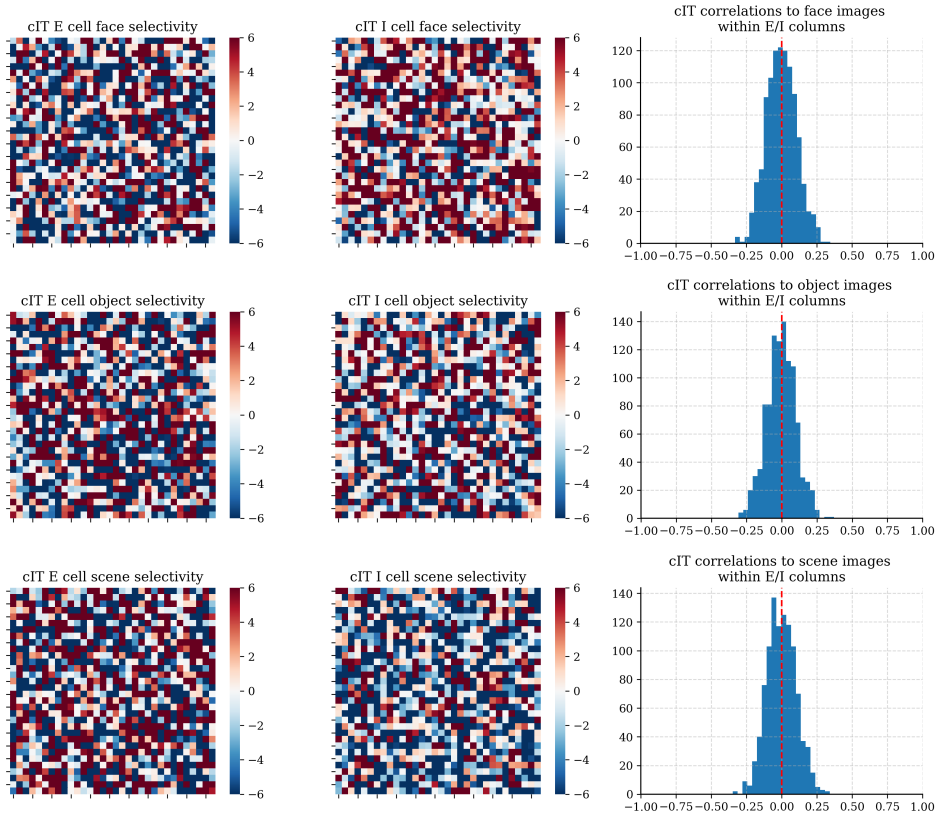


Figure S12: E/I columnar response selectivity in cIT. The selectivity for faces, objects, and scenes, respectively, is plotted for both E and I cells, and the histogram of correlations between E and I cells over all images of the given domain is computed over all locations on the grid, revealing dominantly positive correlations.

9 Assessing representations across the IT hierarchy

A well-known feature of the primate ventral visual stream generally and inferotemporal cortex in particular is increasing view invariance in later vs. earlier regions of the anatomical hierarchy (e.g. Bao et al., 2020). We first asked whether the main ITN model demonstrates this relationship within its IT hierarchy of pIT, cIT, and aIT. To do so, we examined responses to the stimuli of (Bao et al., 2020), containing 51 objects at 24 different viewpoints. As the specific orientation parameters were not available, we computed binary RDMs for each attribute (superordinate and basic category and orientation) using the Hamming distance to determine whether each attribute was identical or different for each image. Additionally, we computed an image-level RDM using the Euclidean distance of images in pixel space. Lastly, we computed RDMs for each layer of IT using the Euclidean distance, and then computed RSA values between model IT areas and various attributes using the Spearman correlation of RDMs (Kriegeskorte et al., 2008). To facilitate plotting the different attributes on the same plot, for each attribute, we normalized the RSA across areas as the % of maximum RSA. Additionally, we computed multivariate decoding accuracy for the three categorical attributes The results, shown in Figure S13 demonstrate that it does not; rather, while the representational similarity seems to imply a *decrease* in view invariance, decoding of category attributes is much flatter, indicating that information is neither lost nor gained moving from pIT to aIT. To place these results in context, we additionally examined an ITN model with a CORnet-Z encoder (Kubilius et al., 2018) and a 2-area IT hierarchy, analyzing each convolutional layer of the encoder along with the IT layers. In this network, we found that view invariance for category attributes (measured with both RSA and decoding accuracy) increased in each stage of the hierarchy with the exception of pIT to aIT (Figure S14). To determine whether the lack of increase in view invariance from pIT to aIT was related to the wiring and sign-based constraints of the ITN, we tested another matched model differing only in that it had no sign constraints and no spatial wiring penalty ($\lambda_w = 0$). This model showed a largely similar pattern (Figure S15); while RSA for the basic-level category attribute increased mildly from pIT to aIT in this model, the decoding accuracy did not, implying that this minor increase in view-invariant RSA was not functionally significant in the same way as the increases in earlier layers which corresponded to increasing category-level information.

Overall, our results highlight that the locally-connected (i.e., fully-connected with a spatial cost) layers used in the current ITN models—combined with current training practices—are not sufficiently powerful to extract increasingly invariant representations beyond a single step. Thus, while stacking a hierarchy of IT layers in an ITN model can provide insight into the constraints on *anatomical* hierarchical topographic organization, further work is needed to extend the ITN to more powerful architectures capable of inducing a corresponding *representational* hierarchy of topographic organization.

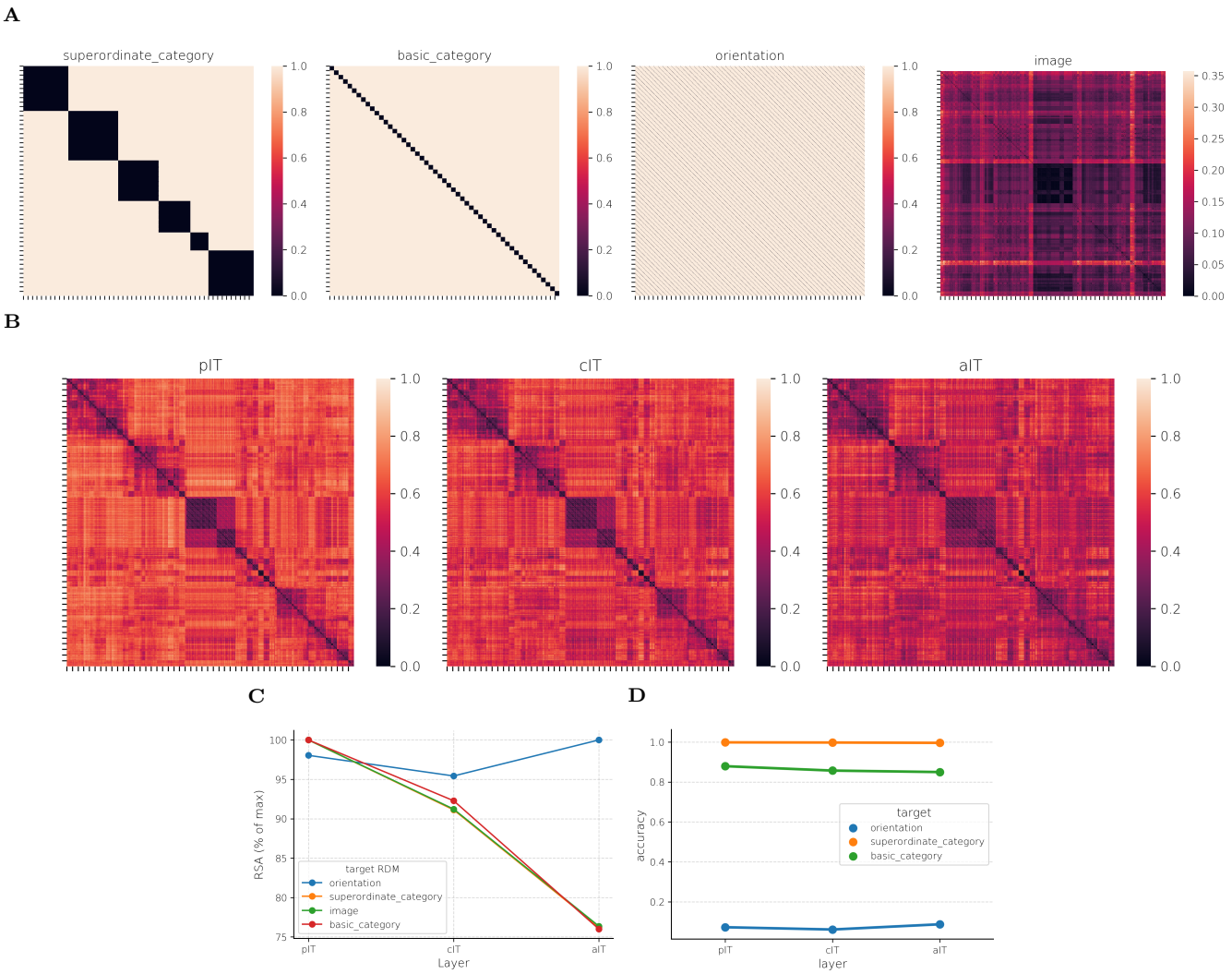


Figure S13: Representational similarity analysis (RSA) at each level of the IT hierarchy in the main ITN model, using the image set of (Bao et al., 2020). **A.** RDMs for each stimulus attribute. For superordinate category, basic category, and orientation, we used a binary notion of distance whereby images with the same stimulus attribute were assigned a distance of 0 and images with different stimulus attributes were assigned a distance of 1 (this was done for orientation as the specific orientation parameters were not available). For the image attribute, Euclidean distance was used. **B.** Euclidean distance RDMs for pIT, cIT, and aIT. **C.** Fractional RSA values (% of max for given attribute over layers), where the RSA values were first computed as 1-r, where r is the Pearson correlation between target and model layer RDMs. **D.** Decoding accuracy for each of 3 categorical attributes using the representations extracted from each layer.

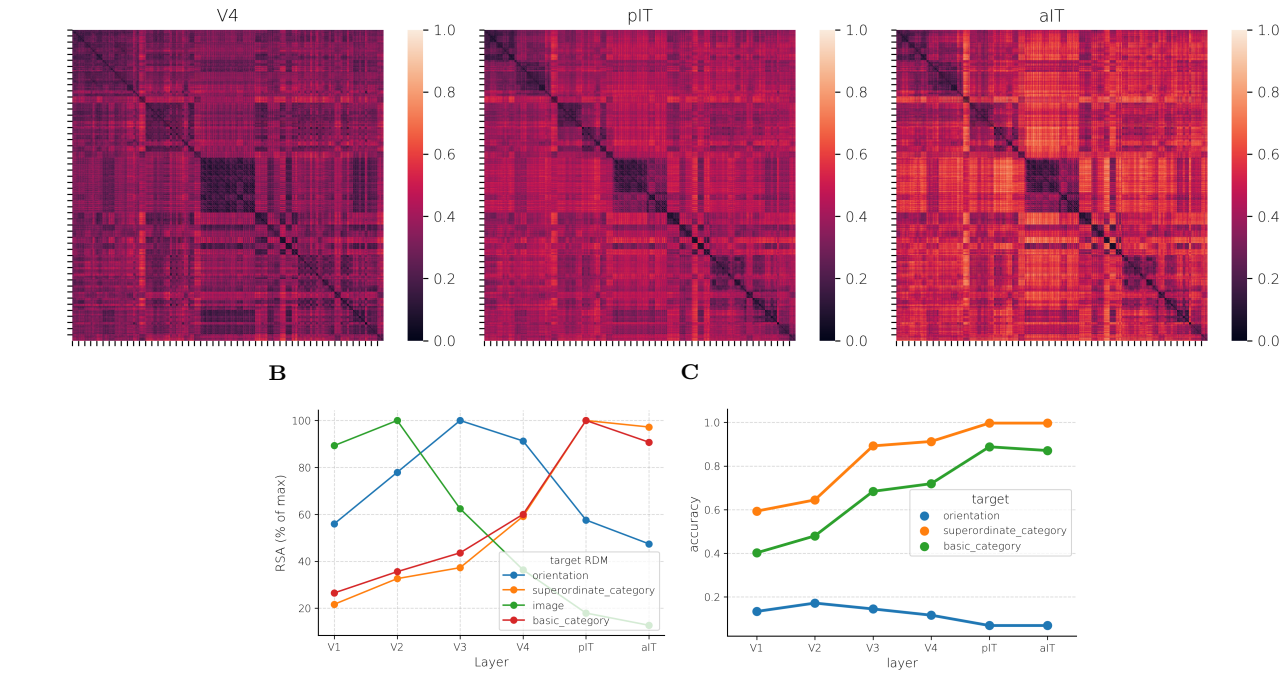


Figure S14: Representational similarity analysis (RSA) in a CORNet-Z encoder version of the model that was trained with rotational data augmentation (max rotation of 120 degrees about vertical), using the image set of (Bao et al., 2020).

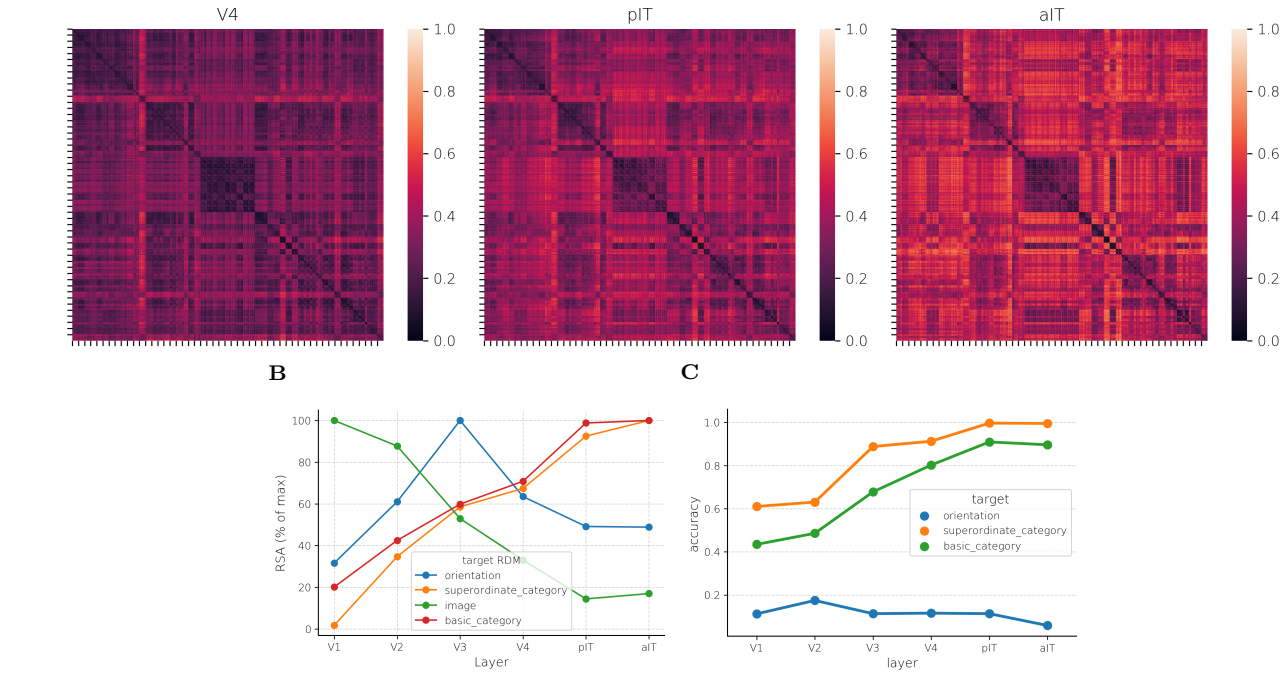


Figure S15: Representational similarity analysis (RSA) in a CORNet-Z encoder RNN ITN model that was trained with rotational data augmentation (max rotation of 120 degrees about vertical) where $\lambda = 0$, using the image set of (Bao et al., 2020).

10 Visualizing the connectivity in a recurrent model constrained to minimize wiring length

Here, we plot a sample of feedforward and lateral connections in the main model used in the paper. Figure 10 plots the learned excitatory feedforward afferents onto cIT neurons, demonstrating localized receptive fields receiving input from the corresponding grid-location of pIT. Figure S17 plots the summed excitatory and inhibitory lateral afferents onto the same cIT neurons, demonstrating localized but larger and largely inhibitory receptive fields. Together, the feedforward local excitatory RFs and lateral primarily inhibitory and wider RFs may encourage the development of smooth topographic organization similar to more classic models (e.g. Kohonen, 1982; Swindale, 1982).

Figure S16: Excitatory feedforward afferents in cIT excitatory neurons. The weights of every other column and row of neurons in the 32x32 grid are plotted, each of which is itself a 32x32 matrix contained within the grid lines.

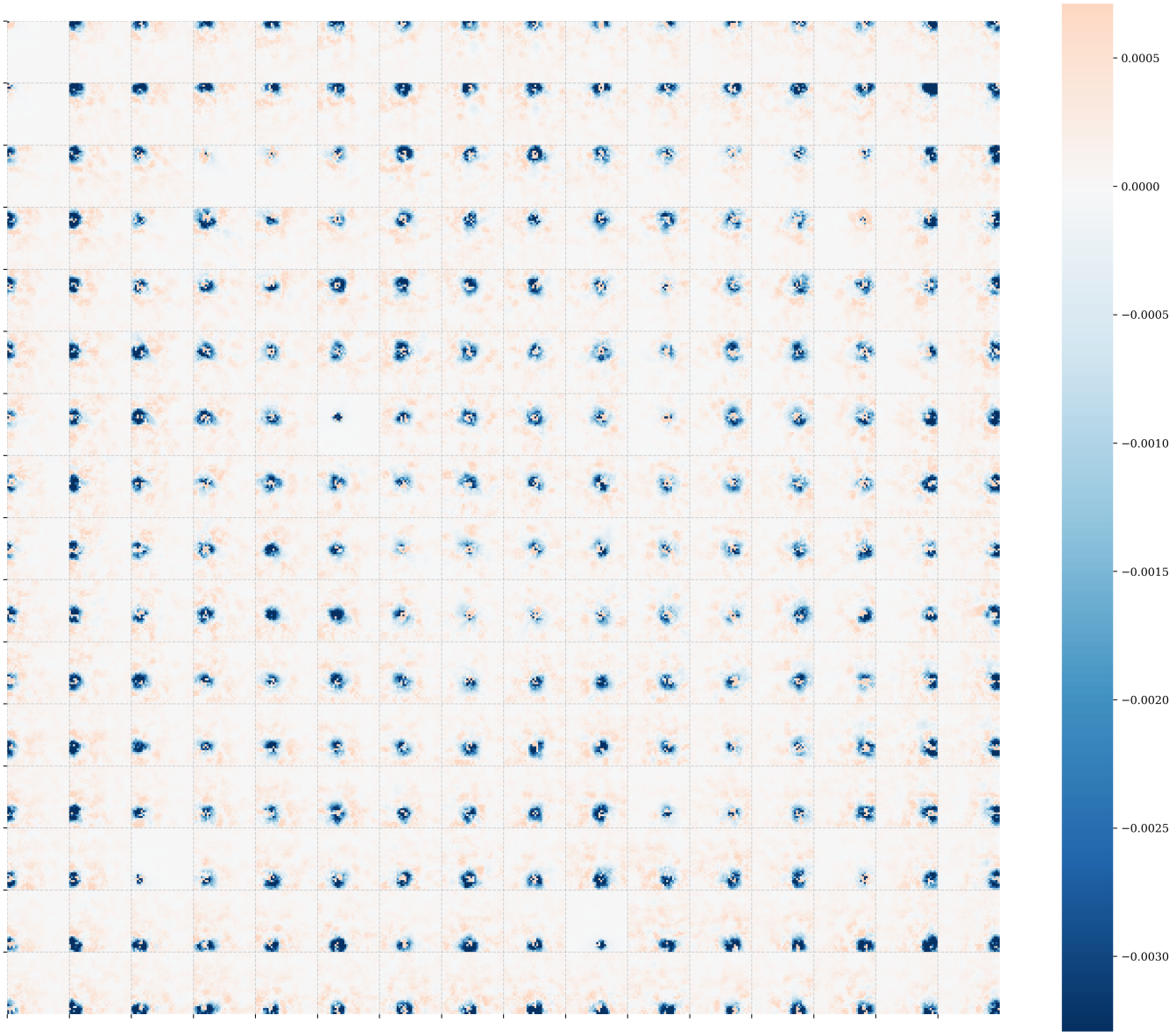


Figure S17: Summed excitatory and inhibitory lateral afferents in cIT excitatory neurons. The weights of every other column and row of neurons in the 32x32 grid are plotted, each of which is itself a 32x32 matrix contained within the grid lines.

11 Analyzing a model trained only on object categories in ImageNet

Concurrent modeling work has suggested that the more general feature space learned through training on ImageNet may be better suited to account for fMRI representational similarity (Prince and Konkle, 2020) and better predicts primate IT face patch activity (Chang et al., 2021) compared to the specialized face and scene feature spaces learned through training on VGGFace2 and Places365, respectively. Moreover, such networks have also been shown to develop face selectivity (Lee et al., 2020). However, previous computational work has also demonstrated that ImageNet-trained DCNN models provide a poor account of human face recognition behavior, both in terms of performance (Blauch et al., 2021) and representational similarity (Dobs et al., 2021). Thus, comparing the pros and cons of networks trained on different image sets remains an important task for research. Here, we investigate a model trained on ImageNet object categories.

Intriguingly, this network produces selectivity for each domain, clustered information, and generic distance-dependent response correlations (Figures S18 and S19), as in the main model trained on all 3 domains. However, we find that local populations of face selective units provide a much weaker amount of information for decoding faces (Figure S18B.)—reaching only approximately 20% decoding—compared to local populations of face-selective units in a network that has been trained on all 3 domains (Figure 1, reaching approximately 75% decoding), in line with previous work (Blauch et al., 2021).

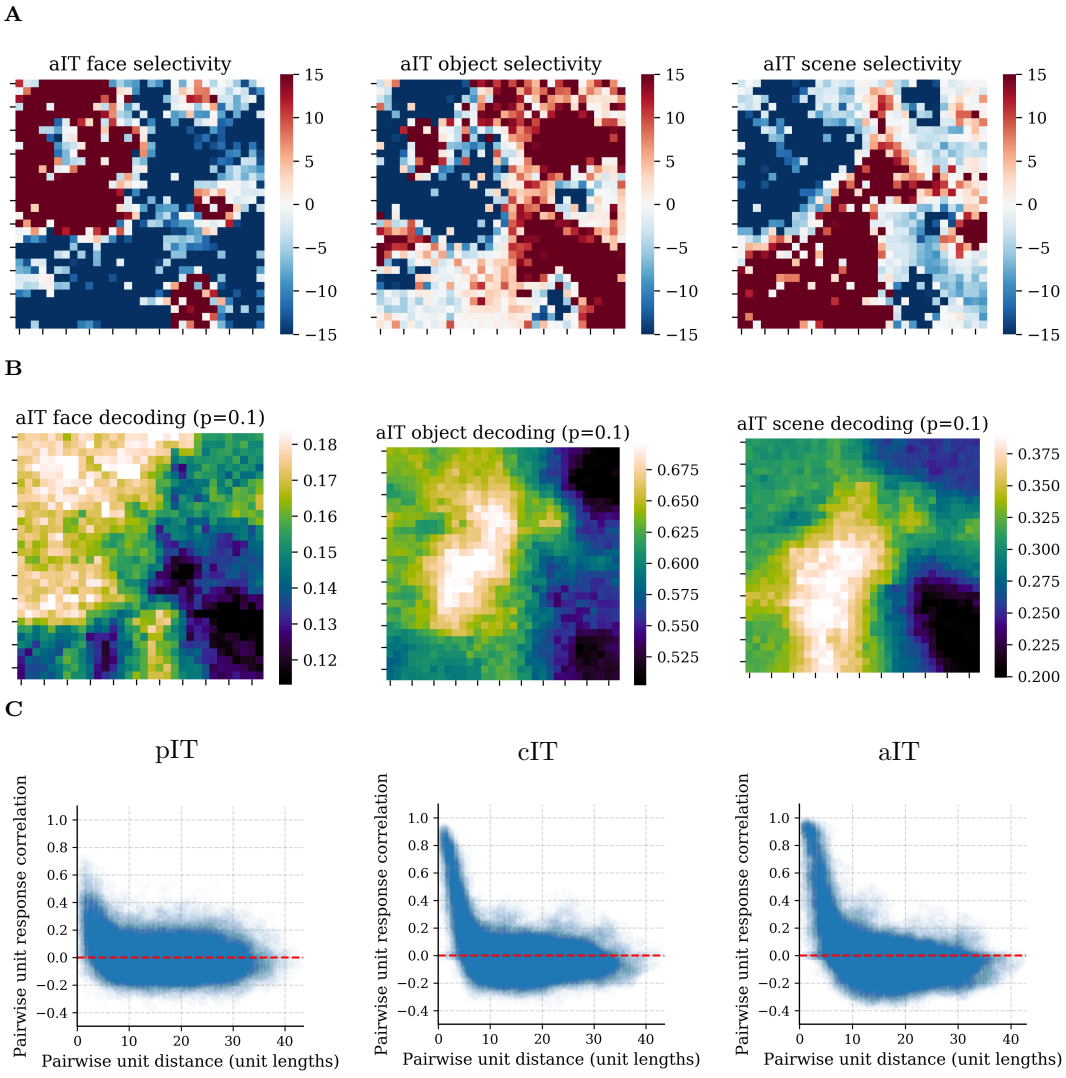
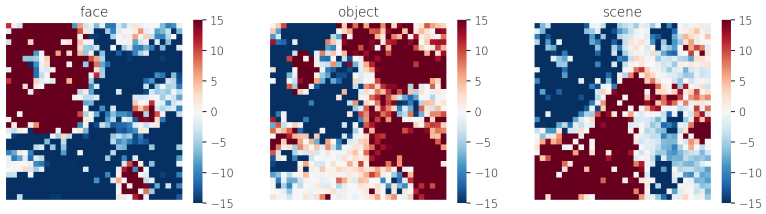
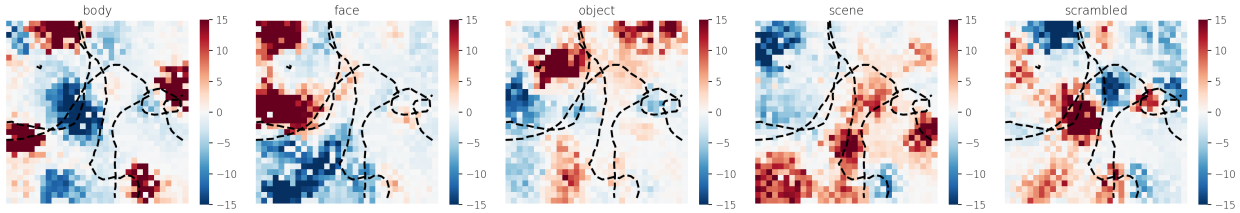


Figure S18: Results for the main architecture, trained only on object categories. Note that the scales for searchlight plots (in **B**) are allowed to vary by domain, since the network has very different levels of information for each domain, in particular very weak information for discriminating face identities. Domain-level lesion analyses were not performed as the model was trained only on a single domain.

A Validation images from trained face, object, and scene domains



B Validation images from trained face, object, and scene domains



C

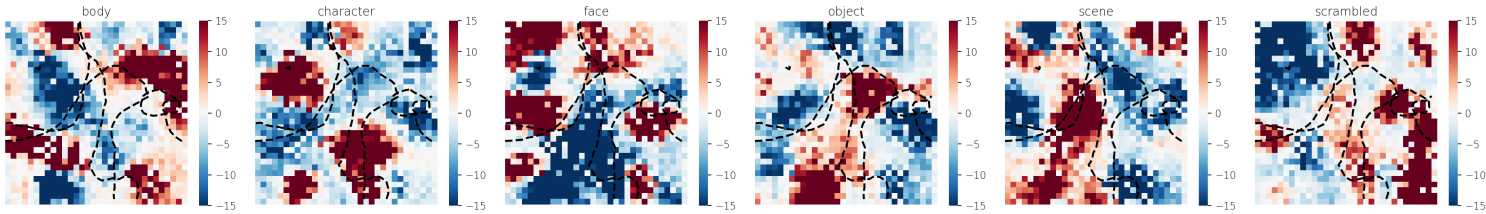


Figure S19: Selectivity to multiple stimulus sets for the ImageNet-trained model. **A.** validation images used throughout the main text, for reference. **B.** A localizer set from the Konkle lab. **C.** The *fLoc* stimulus set from the Grill-Spector lab (Stigliani et al., 2015). Contours in **B** and **C** show outlines of significant smoothed domain selectivity ($p<0.001$; smoothing done by averaging using 5% nearest units).

12 Further analysis of principal components of activation

In the main paper, we examined the principal components to activations using un-trained images from the trained object, face, and scene domains. Here, we examined a set of computer generated object stimuli to ask whether the main ITN reproduced the more precise details of organization discovered by Bao et. al (Bao et al., 2020)—specifically, clustering of stubby, spiky, face, and body/animal stimuli in 4 quadrants of a PC1-PC2 space. We acquired the stimuli of Bao et. al and generated a PC1-PC2 space; following them, we colored the 100 images that maximally activated each network in electrophysiological recordings. We found strong clustering of body, face, and stubby stimuli, and weaker clustering of spiky stimuli (Figure S20D.,top left). However, the weights of the PCs were less cleanly topographically organized (Figure S20D.,top right) than those discovered when the PC space was constructed using within-distribution (but unseen) images from the trained domains (Figure S20A., top). Moreover, the selectivity to the localizer stimuli used by Bao et. al to localize these regions in macaques was topographically organized to some extent – in particular for the face and body clusters – but was more diffuse for spiky and stubby clusters (Figure S20D, bottom right). An ITN model using a weaker spatial penalty ($\lambda = 0.1$) showed somewhat less diffuse patchy organization for the Bao stimuli (Figure S22), implying that a weaker wiring cost might yield topographic organization that is more robust to distributional differences between image sets.

We next examined the principal components of activation in an ImageNet-trained version of the main ITN architecture. Like the main model, this network yielded clustered selectivity for each domain (Figure S18), and topographically organized within-domain attributes (Figure S20E). In contrast to the main model, the object-constructed PCs now organized across the whole map, with the animacy distinction corresponding to the spatial X axis of aIT. Face gender was less well discriminated, but intriguingly, face selectivity and PC weight were nonetheless topographically organized and localized with greater mass in the inanimate—rather than the animate— part of the map, but not exclusively. This suggests that faces may share similarity with inanimate objects—possibly in terms of curvilinear features (Yue et al., 2020)—as well as animate objects–such as animals where faces are a salient aspect of visual appearance and human recognition behavior. Notably, this is in contrast to empirical data from human fMRI in which face selectivity systematically appears within animate-preferring zones of ventral temporal cortex (Konkle and Caramazza, 2013).

Last, we asked whether this ImageNet-trained ITN model might better account for the primate organization of the stimuli from Bao et. al. As shown in Figure S20H., this model yielded somewhat more smooth PCs for representing the object stimuli, and led to somewhat more cleanly topographically organized regions selective for each category (again, stubby images elicited the most diffuse responses). A version of this model with a weaker wiring penalty ($\lambda_w = 0.1$) revealed even neater animacy organization (Figure S22E), and possibly neater cluster-level organization for the Bao stimuli, especially when viewing selectivity of locally smoothed activations more akin to the BOLD signal used in fMRI (Figure S22F). This suggests that the more general statistics of ImageNet images may lead to a somewhat more biologically-plausible topographic organization in ImageNet-trained ITN models. However, a more detailed quantitative comparison of different ITN models with neural data is a large effort beyond the scope of this work, and will be undertaken in future work.

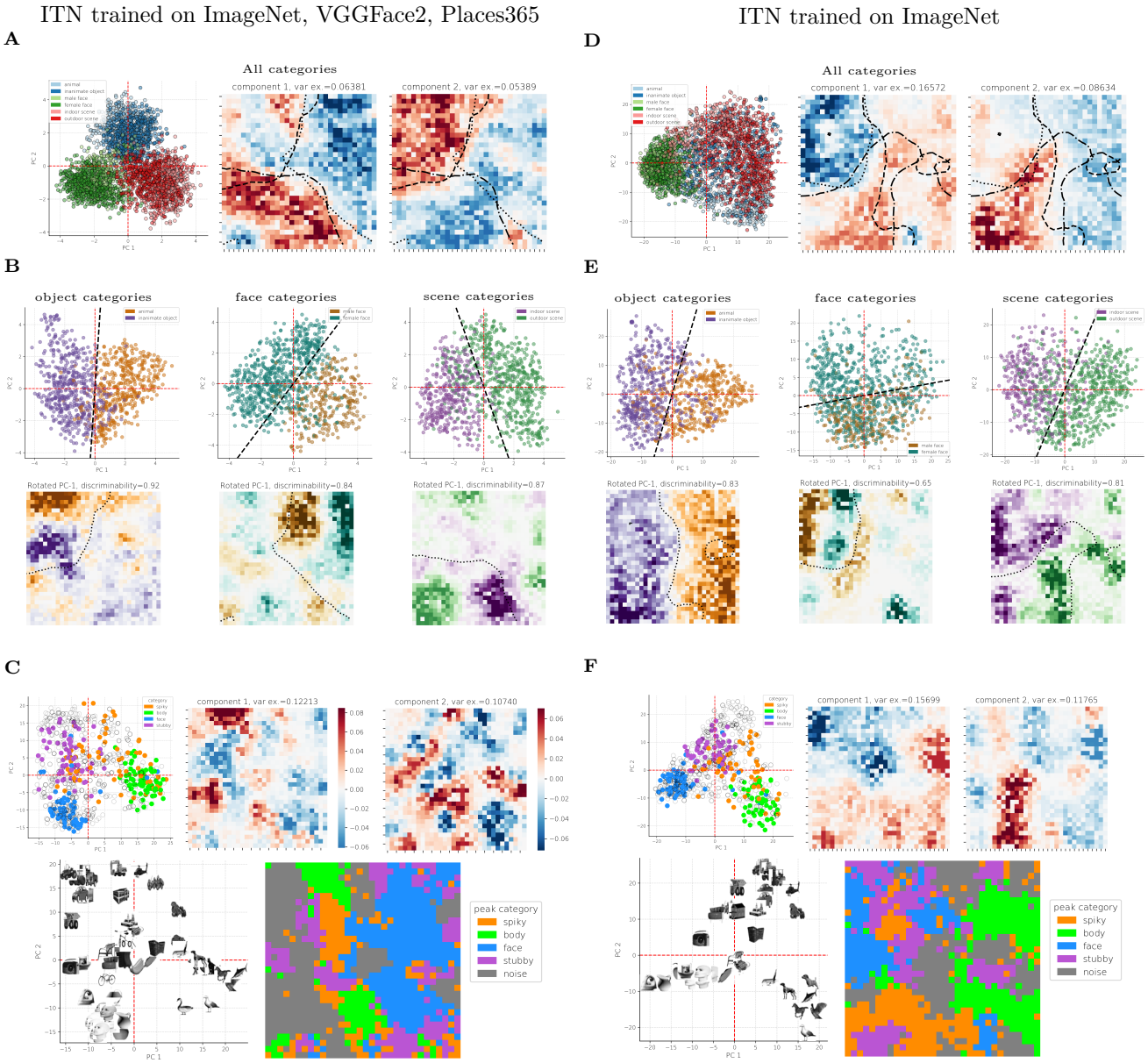


Figure S20: Principal components analysis of activations. **A-C** Main model. **D-F** ImageNet trained version of the main architecture. **A.** and **D.** plot the domain selectivity. **B.** and **E.** plot the PC1-PC2 space and PC1 and PC2 component weights across images from all three domains (top), the PC1-PC2 space and a rotated component weight that maximized the discriminability of images according to a given sub-domain attribute (gender for faces, animacy for objects, and indoor/outdoor for scenes). **C.** and **F.** show a principal components analysis of images from Bao et. al (Bao et al., 2020). In the top-left plot, the 100 images eliciting highest activation in each macaque "network" (see Bao et. al) are colored by network. In the lower panel, we plot selectivity for the localizer stimuli used to discover these networks, where the color indicates the preferred category.

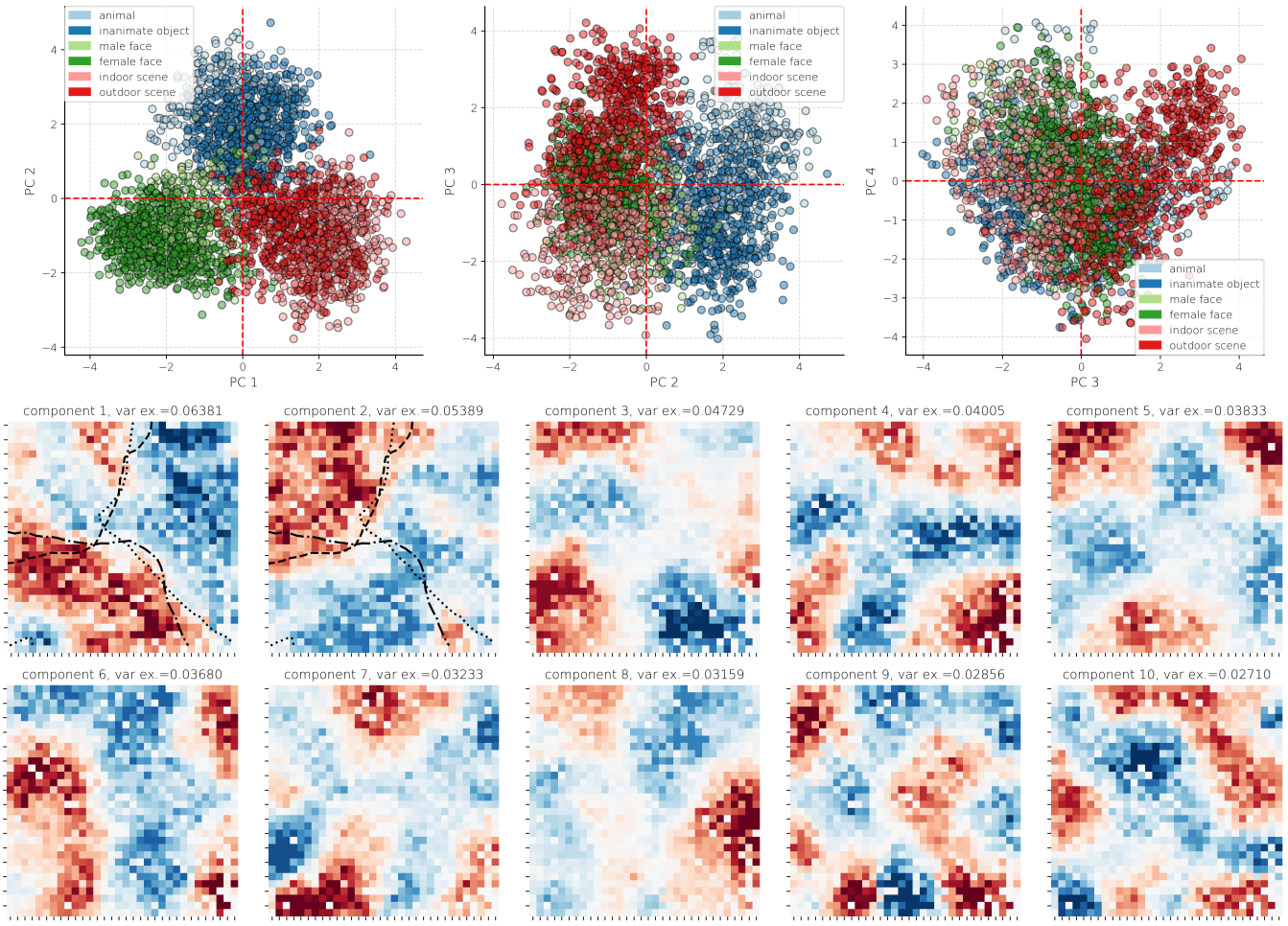


Figure S21: Visualizing higher PCs in the main model.

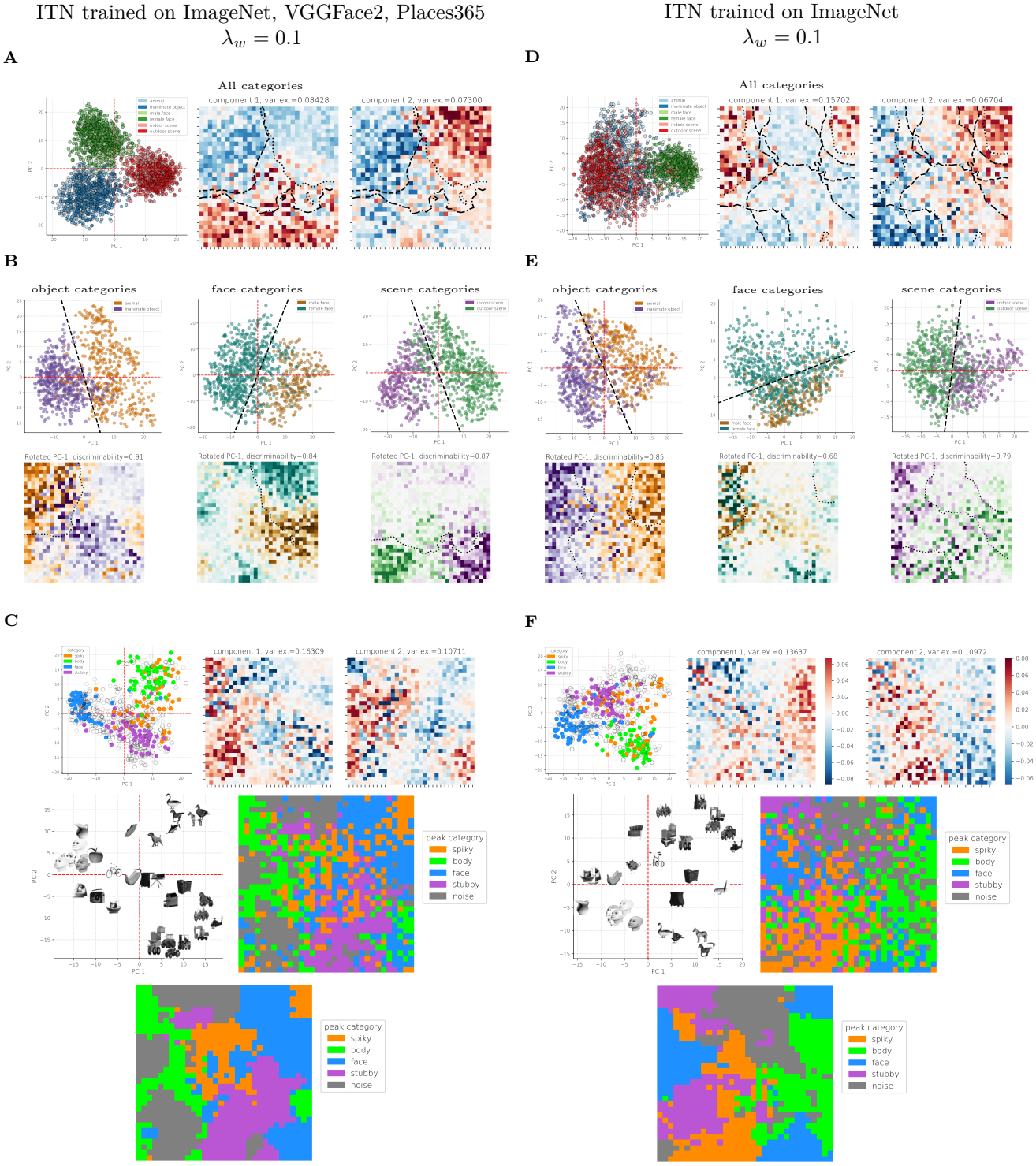


Figure S22: Principal components analysis of activations in models with a weaker wiring cost. **A-C** Main model. **D-F** ImageNet trained version of the main architecture. **A.** and **D.** plot the domain selectivity. **B.** and **E.** plot the PC1-PC2 space and PC1 and PC2 component weights across images from all three domains (top), the PC1-PC2 space and a rotated component weight that maximized the discriminability of images according to a given sub-domain attribute (gender for faces, animacy for objects, and indoor/outdoor for scenes). **C.** and **F.** show a principal components analysis of images from Bao et. al (Bao et al., 2020). In the top-left plot, the 100 images eliciting highest activation in each macaque "network" (see Bao et. al) are colored by network. In the lower panel, we plot selectivity for the localizer stimuli used to discover these networks, where the color indicates the preferred category. The lowest panel then plots selectivity using locally smoothed activations (10-nearest-neighbor averaging kernel).

13 Robustness of results

To verify the robustness of our results, we ran multiple versions of the main ITN model using different random seeds controlling the random weight initialization and training stimulus presentation order, both of which can in theory effect the emergent functional organization. We plot these results in Supplementary Figure S23.

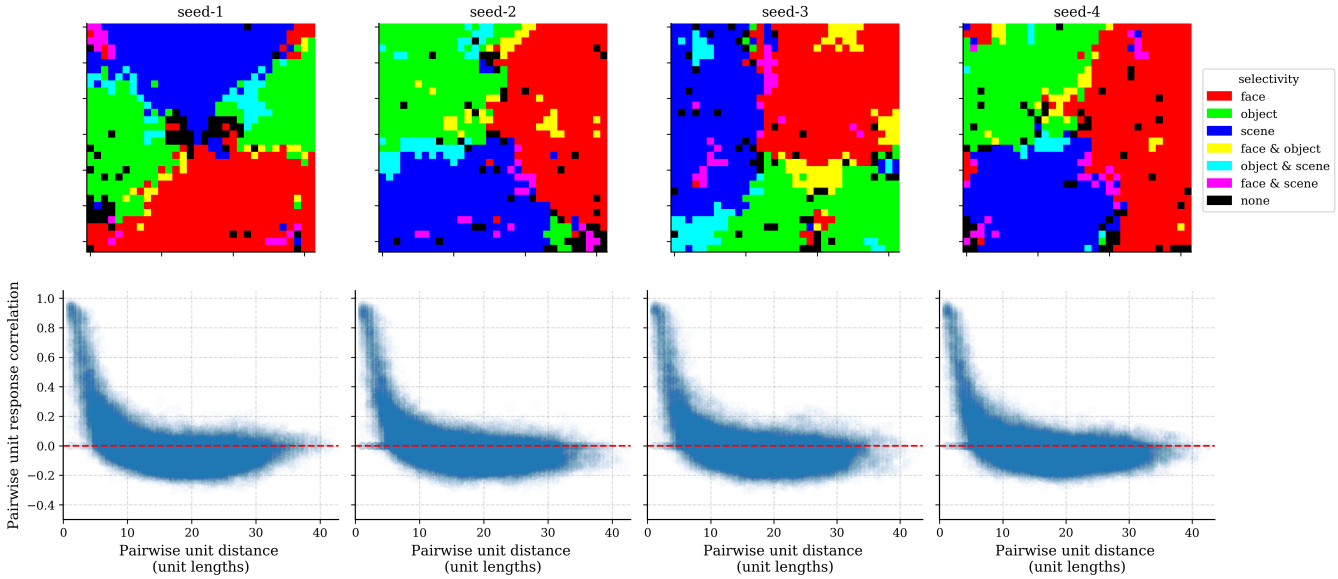


Figure S23: Robustness of results. Top: domain-selective topography in aIT. Bottom: generic distance-dependent response correlations in aIT. Note that seed-2 corresponds to the main model used in the paper.

14 Detailed results for an intermediate model, with strictly excitatory feedforward connections but no separation of E and I (EFF RNN)

In the main paper, we presented an ITN variant that utilizes excitatory feedforward connectivity, layer normalization, and recurrent lateral connectivity. In Supplementary Figure S25, we plot a larger range of results for this model, including domain selectivity, within-domain information, lesion deficits, and generic topography, along with the full λ_w tuning analyses used to select the optimal model presented in the main text.

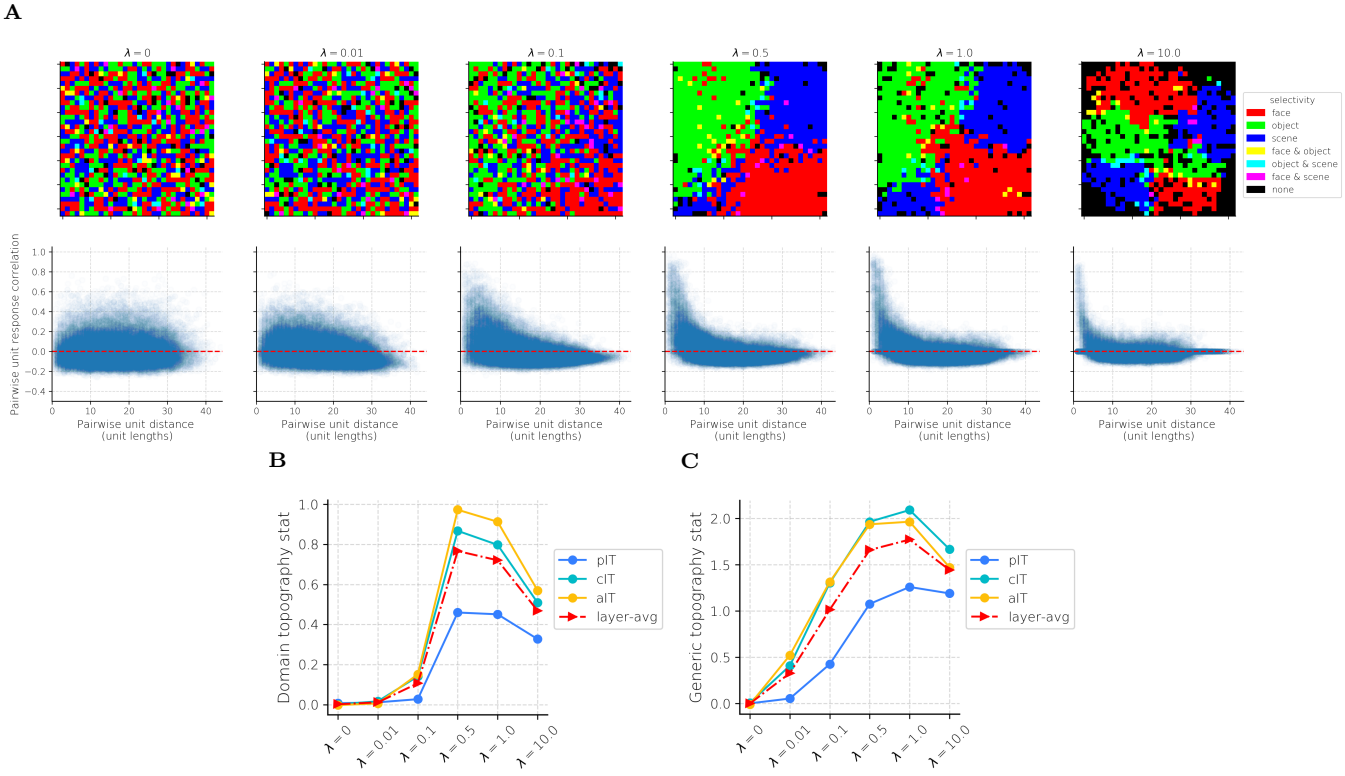


Figure S24: Domain-level and generic topography as a function of λ_w in the intermediate model with strictly excitatory feedforward connections but no separation of E and I.

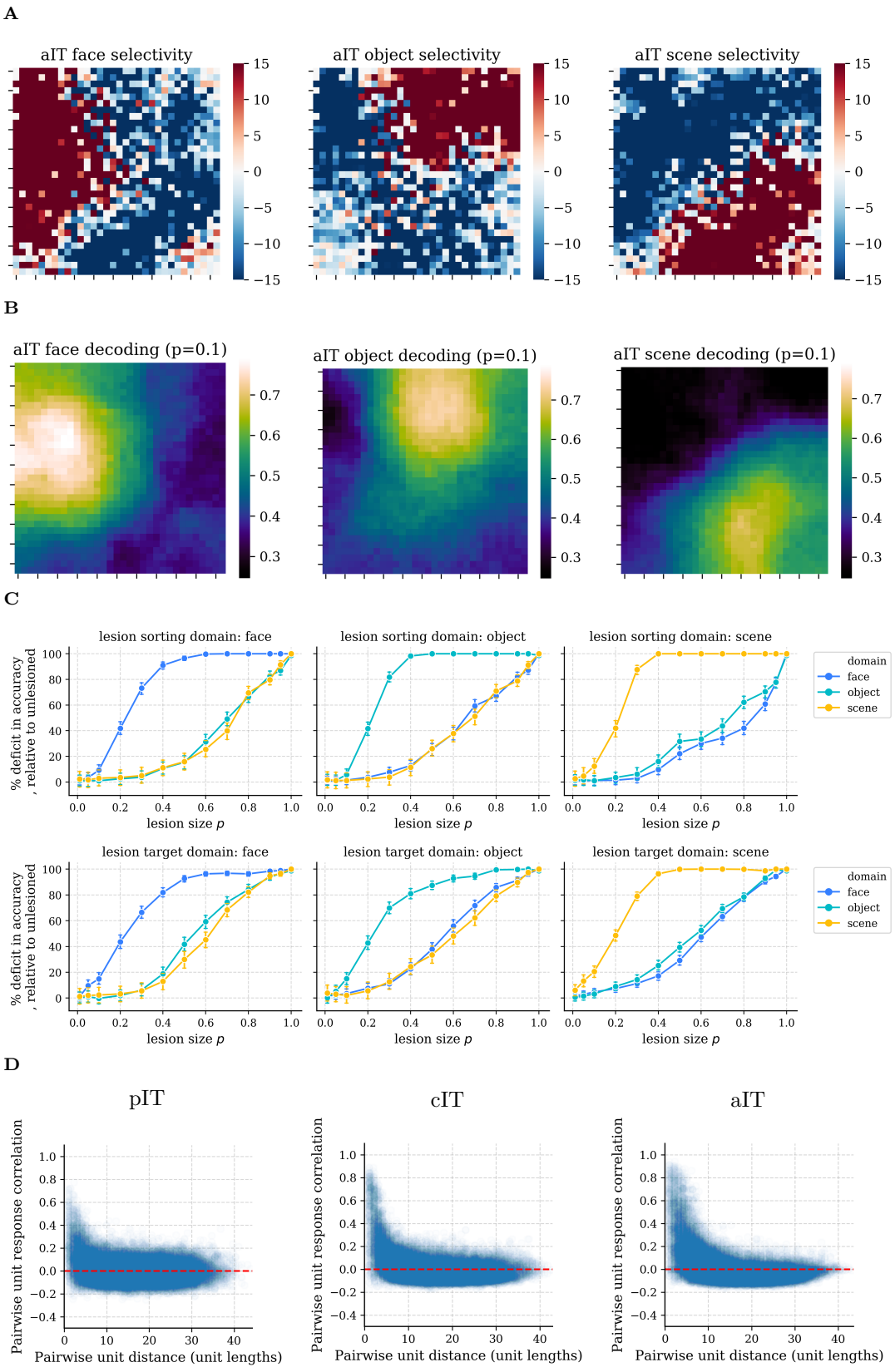


Figure S25: A simplified model containing local, excitatory feedforward connectivity, local lateral connectivity (one sheet of units without sign constraints), and global lateral divisive inhibition (layer normalization) exhibits domain-level and generic topographic organization very similar to the more biologically-detailed model.

15 Detailed results for a model with separation of E and I but no restriction on the sign of feedforward connectivity (E/I RNN)

In the main paper, we presented an ITN variant that separates excitation and inhibition into parallel sheets of neurons, but does not restrict feedforward afferents to the E units (E/I RNN). In Supplementary Figure S27, we plot a larger range of results for the E/I RNN model, including domain selectivity, within-domain information, lesion deficits, and generic topography, along with the full λ_w tuning analyses used to select the optimal model presented in the main text.

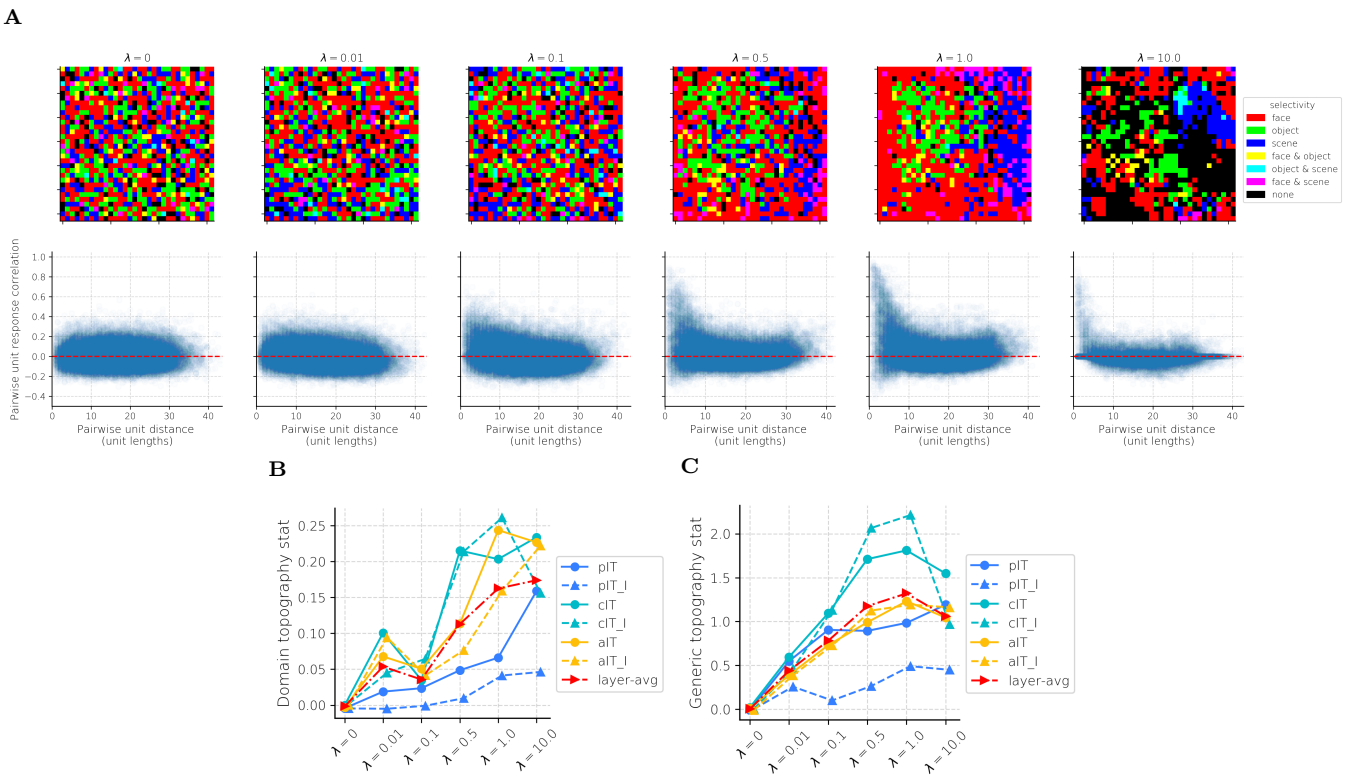


Figure S26: Domain-level and generic topography as a function of λ_w in the E/I FNN model. **A.** Emergent topography. **B.** Quantification of domain-level topography in terms of the dot product between 3-D vectors of domain selectivity, scaled by distance and summed over unit pairs.

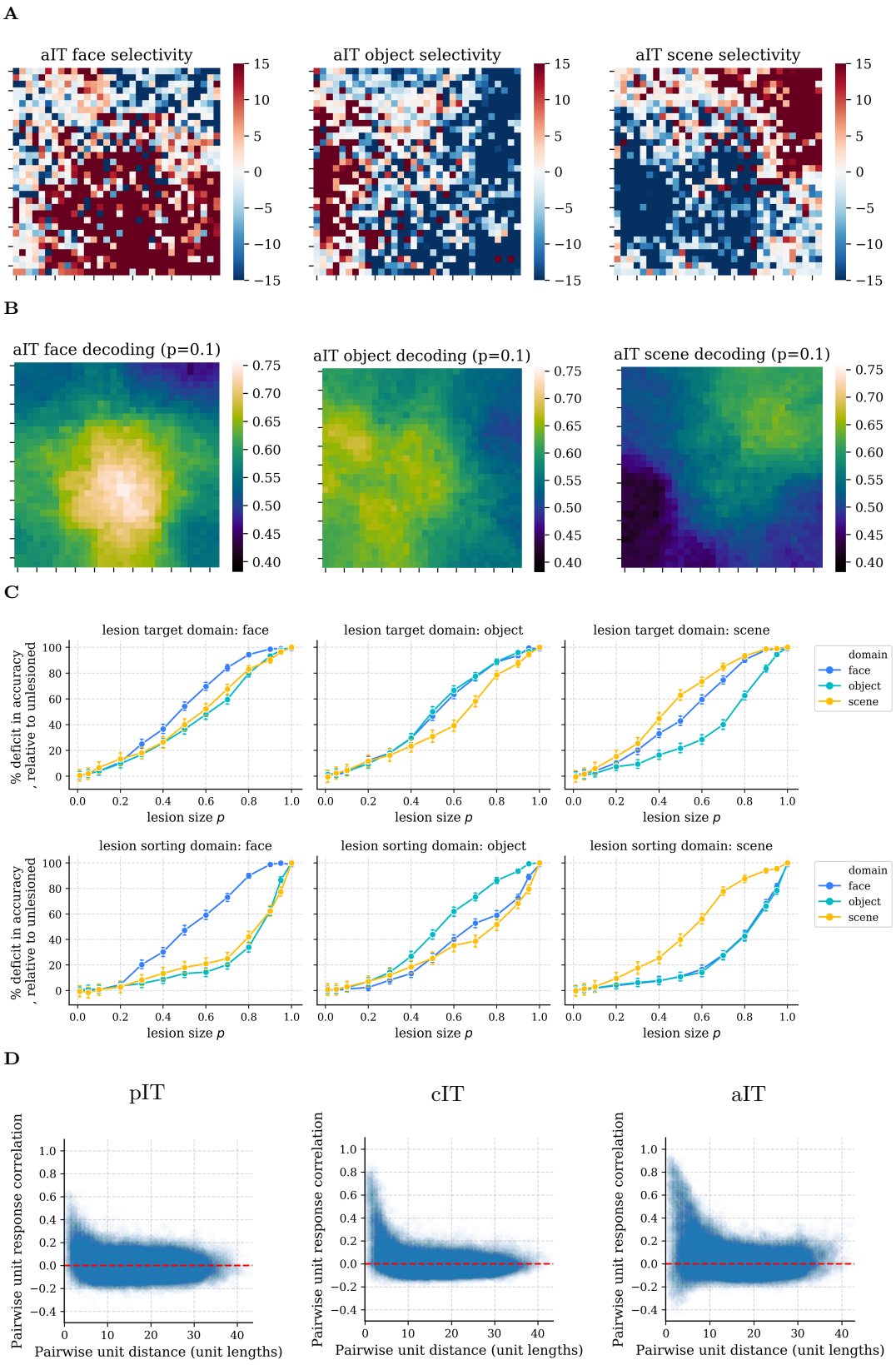


Figure S27: An intermediate model, with separation of E and I but no restriction on the sign of feedforward connectivity, and global lateral divisive inhibition (layer normalization) exhibits domain-level and generic topographic organization very similar to the more biologically-detailed model.

16 Detailed results for a feedforward model, with separation of E and I but no restriction on the sign of feedforward connectivity (E/I FNN)

In the main paper, we presented a feedforward ITN variant that separates excitation and inhibition into parallel sheets of neurons, but does not restrict feedforward afferents to the E units (E/I FNN). In Supplementary Figure S27, we plot a larger range of results for the E/I FNN model, including domain selectivity, within-domain information, lesion deficits, and generic topography, along with the full λ_w tuning analyses used to select the optimal model presented in the main text.

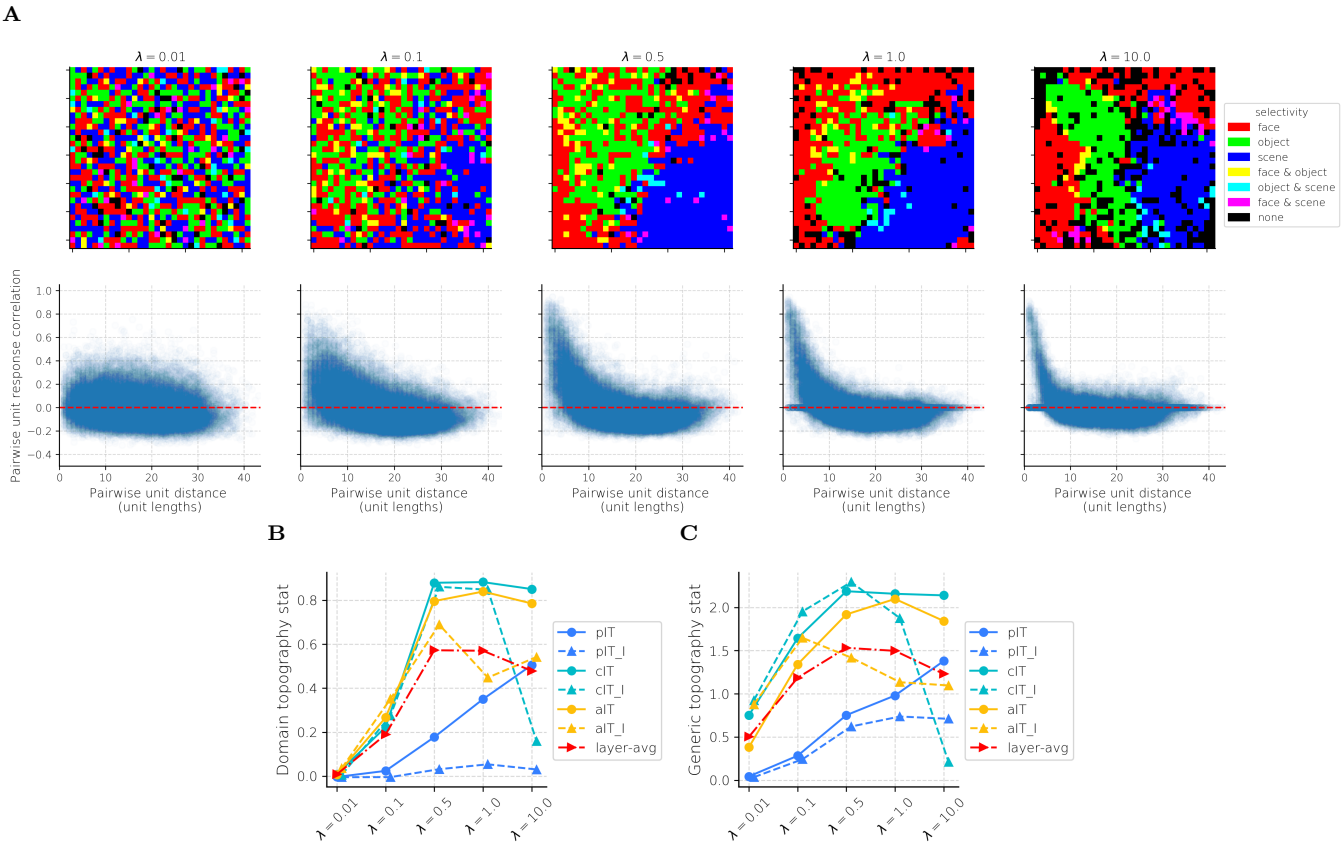


Figure S28: Domain-level and generic topography as a function of λ_w in the E/I FNN model. **A.** Emergent topography. **B.** Quantification of domain-level topography in terms of the dot product between 3-D vectors of domain selectivity, scaled by distance and summed over unit pairs.

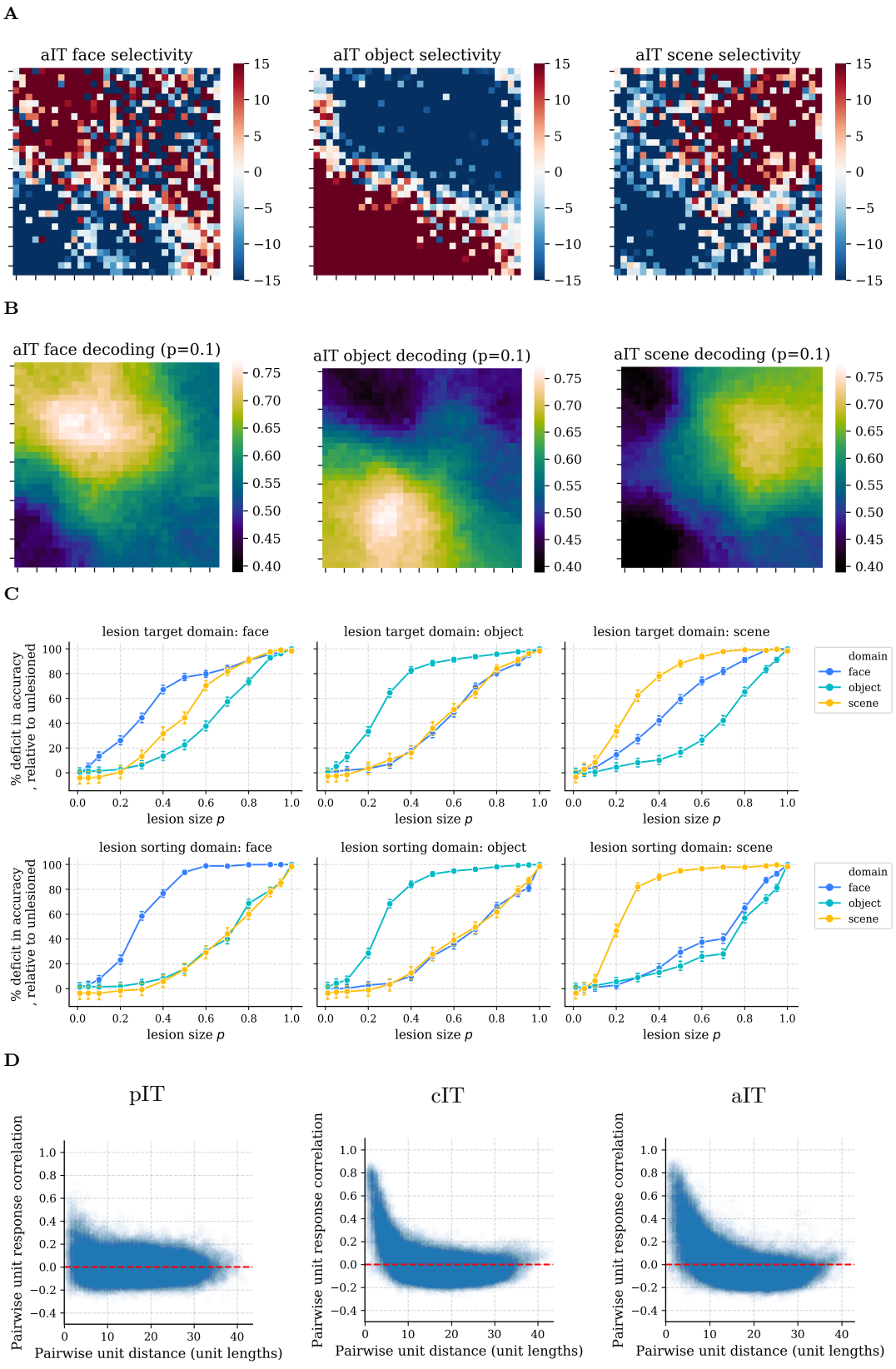


Figure S29: Detailed results for the E/I FNN model.

17 Detailed results for a feedforward model with excitatory-only feedforward connections and no separation of E/I (EFF FNN)

In the main paper, we presented an ITN variant that utilizes excitatory feedforward connectivity and layer normalization without learned lateral connectivity (EFF FNN).

In Supplementary Figure S31, we plot a larger range of results for this model, including domain selectivity, within-domain information, lesion deficits, and generic topography, along with the full λ_w tuning analyses used to select the optimal model presented in the main text.

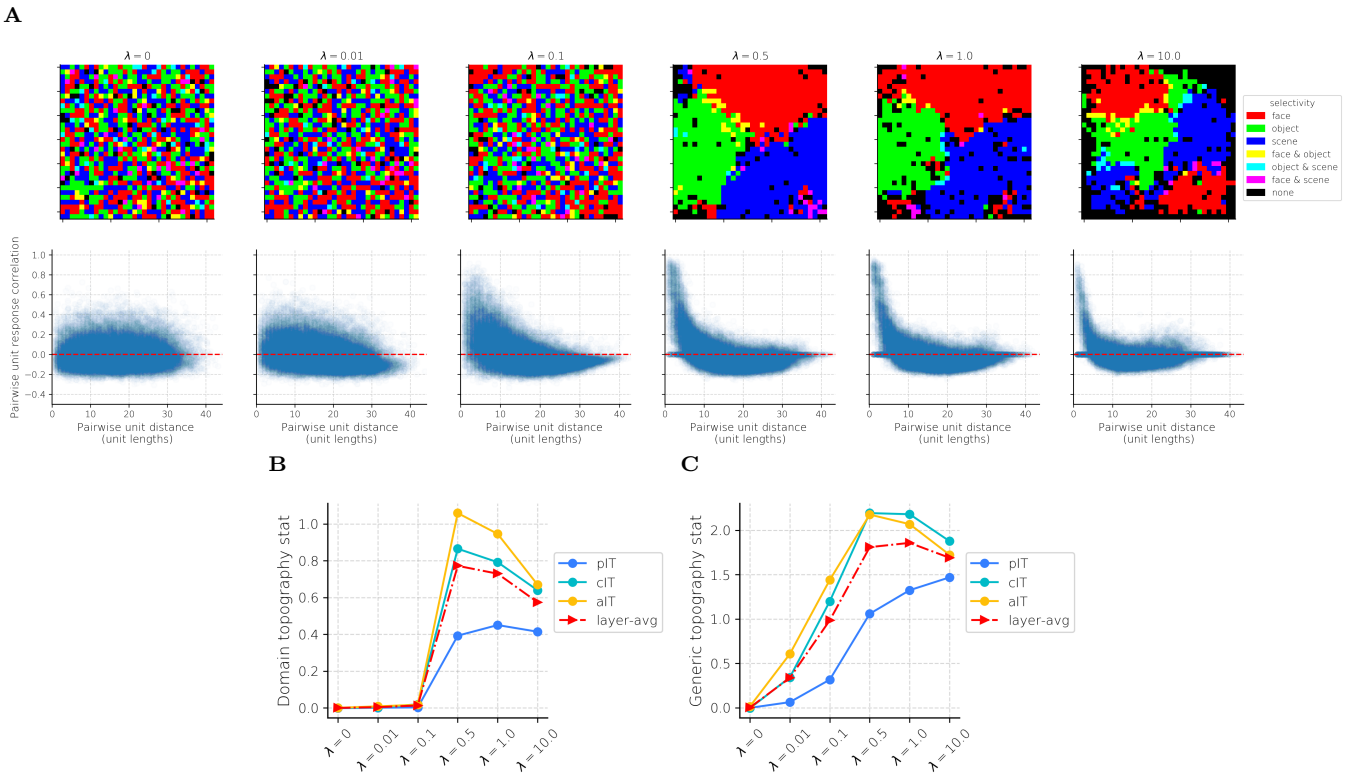


Figure S30: Domain-level and generic topography as a function of λ_w in the simplified feedforward model.

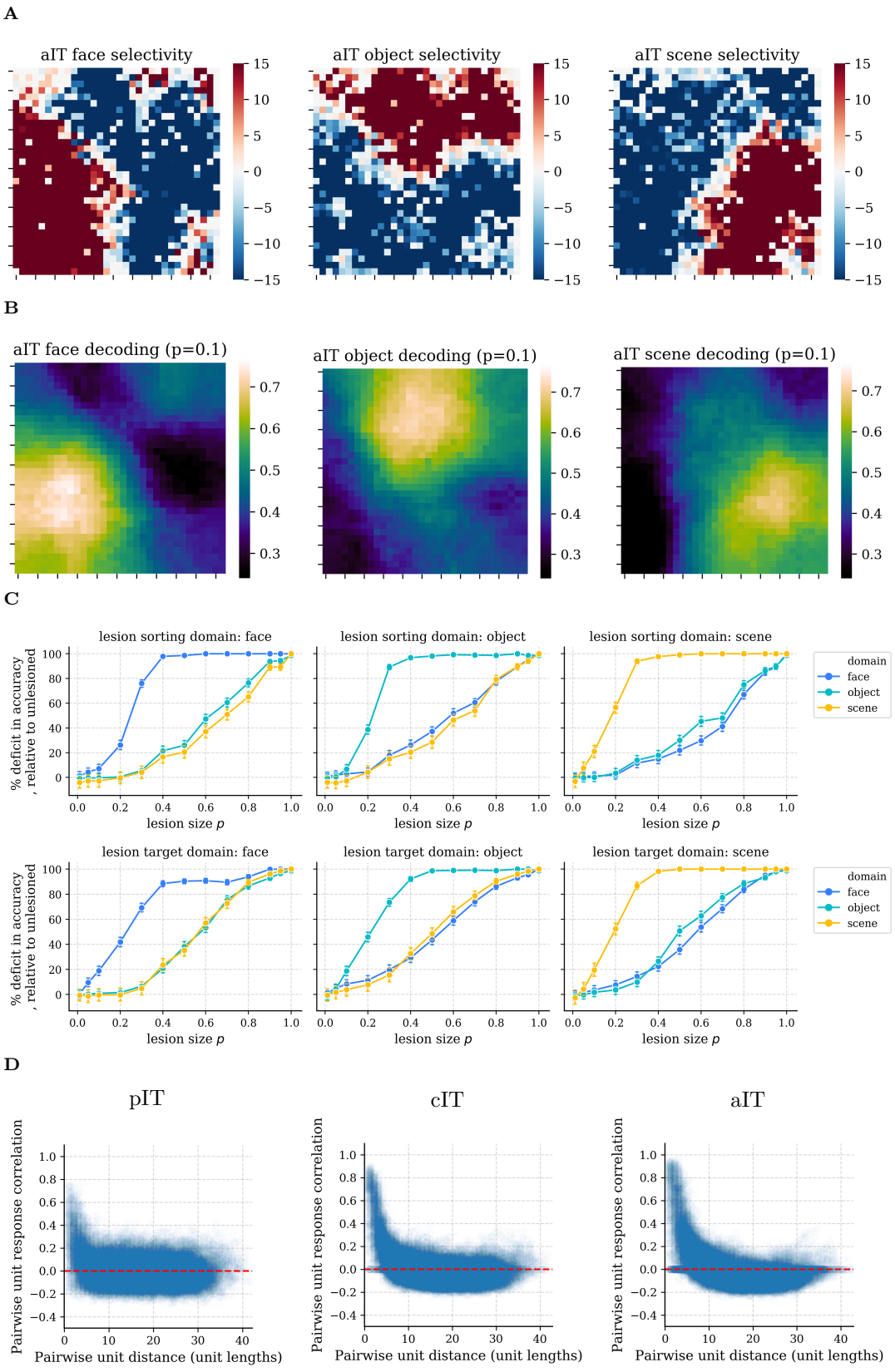


Figure S31: A simplified model containing local, excitatory feedforward connectivity and global lateral divisive inhibition (layer normalization) exhibits domain-level and generic topographic organization very similar to the more biologically-detailed model. See the main text (Figures 1, 3, 4) for more details on the plots.

18 Detailed results for a model with wiring minimization but no sign constraints (RNN)

In the main paper, we presented an ITN variant without sign constraints (RNN).

In Supplementary Figure S33, we plot a larger range of results for this model, including domain selectivity, within-domain information, lesion deficits, and generic topography, , along with the full λ_w tuning analyses used to select the optimal model presented in the main text.

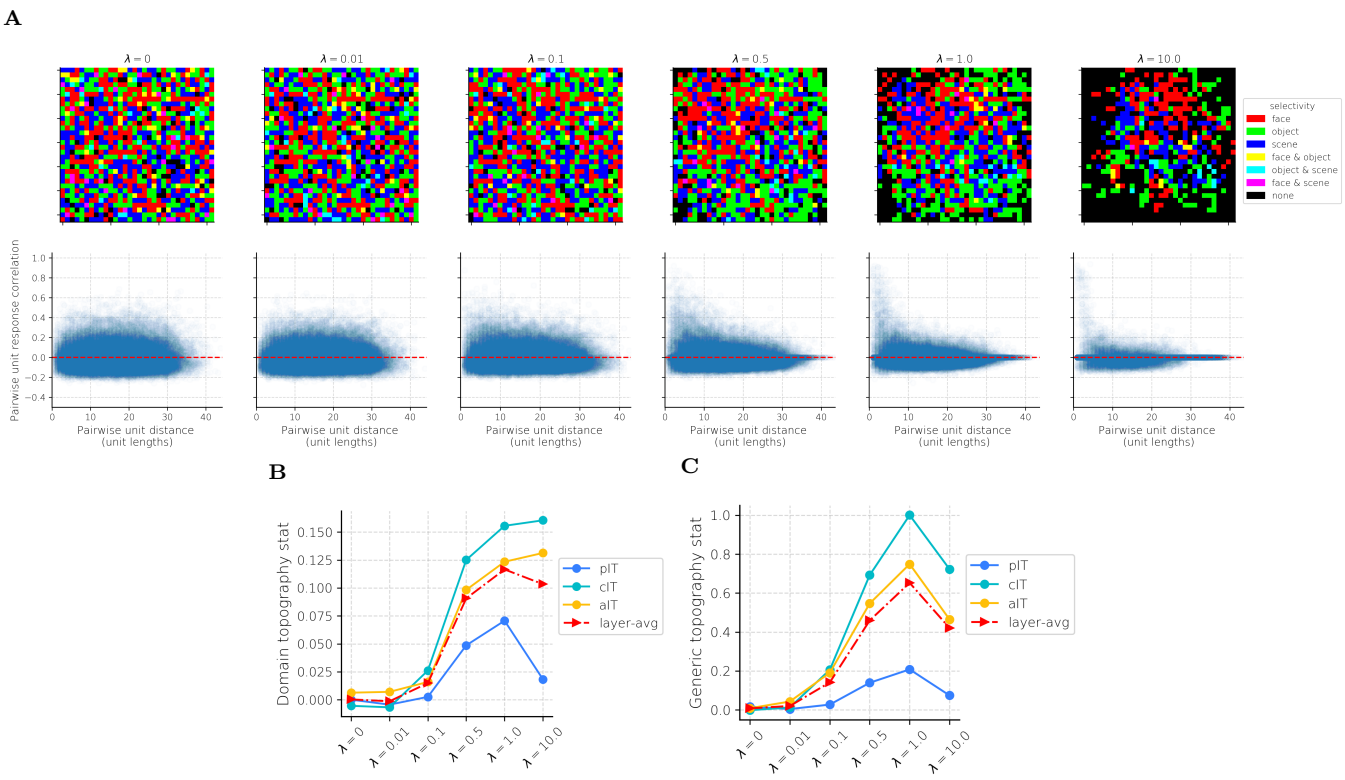


Figure S32: Domain-level and generic topography as a function of λ_w in a model without sign constraints. **A.** Emergent topography. **B.** Quantification of domain-level topography in terms of the dot product between 3-D vectors of domain selectivity, scaled by distance and summed over unit pairs.

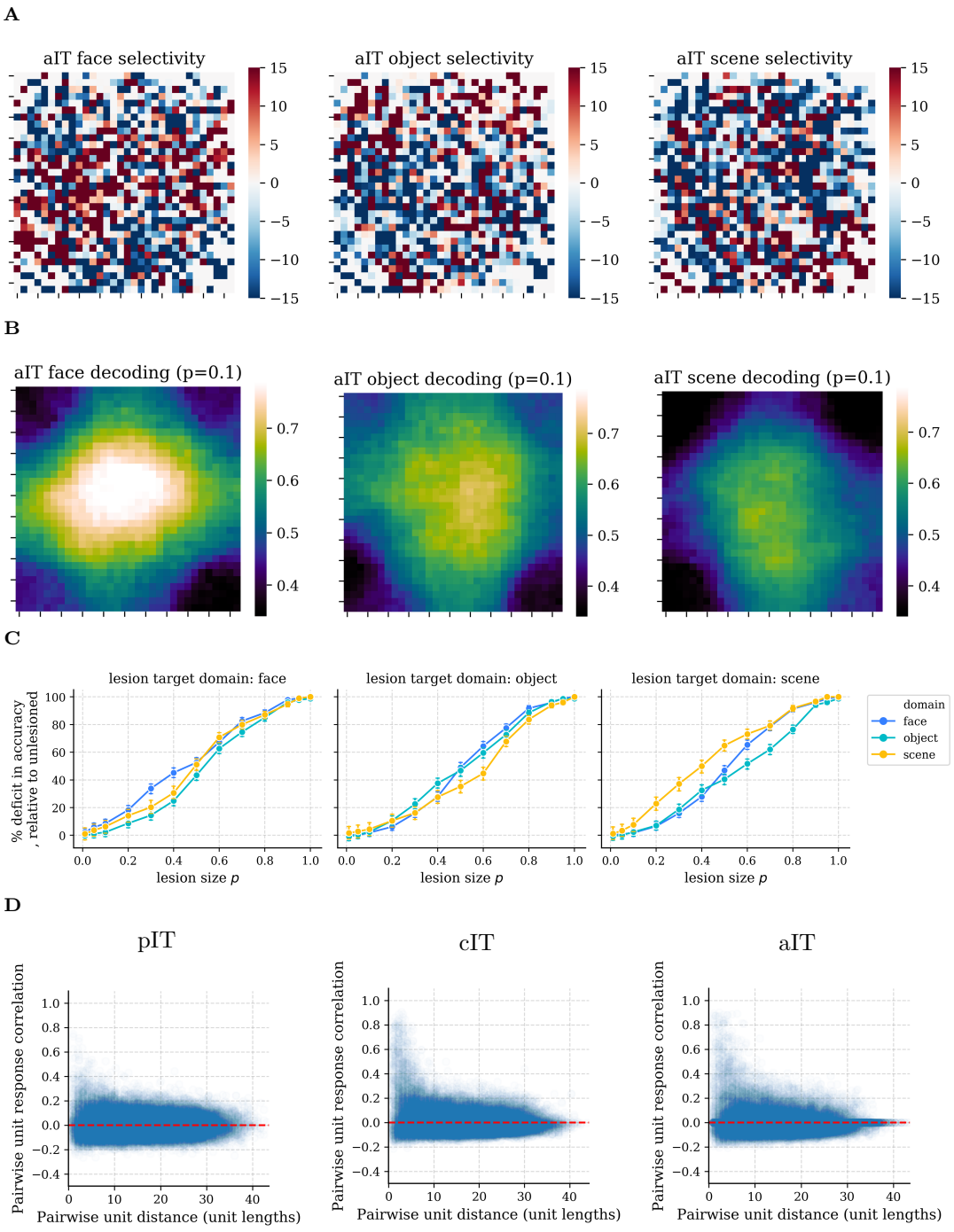


Figure S33: Results for a model without sign constraints (RNN).

19 Detailed results for a feedforward model with wiring minimization but no sign constraints (FNN)

In the main paper, we presented an ITN variant without sign constraints (RNN).

In Supplementary Figure S33, we plot a larger range of results for this model, including domain selectivity, within-domain information, lesion deficits, and generic topography, , along with the full λ_w tuning analyses used to select the optimal model presented in the main text.

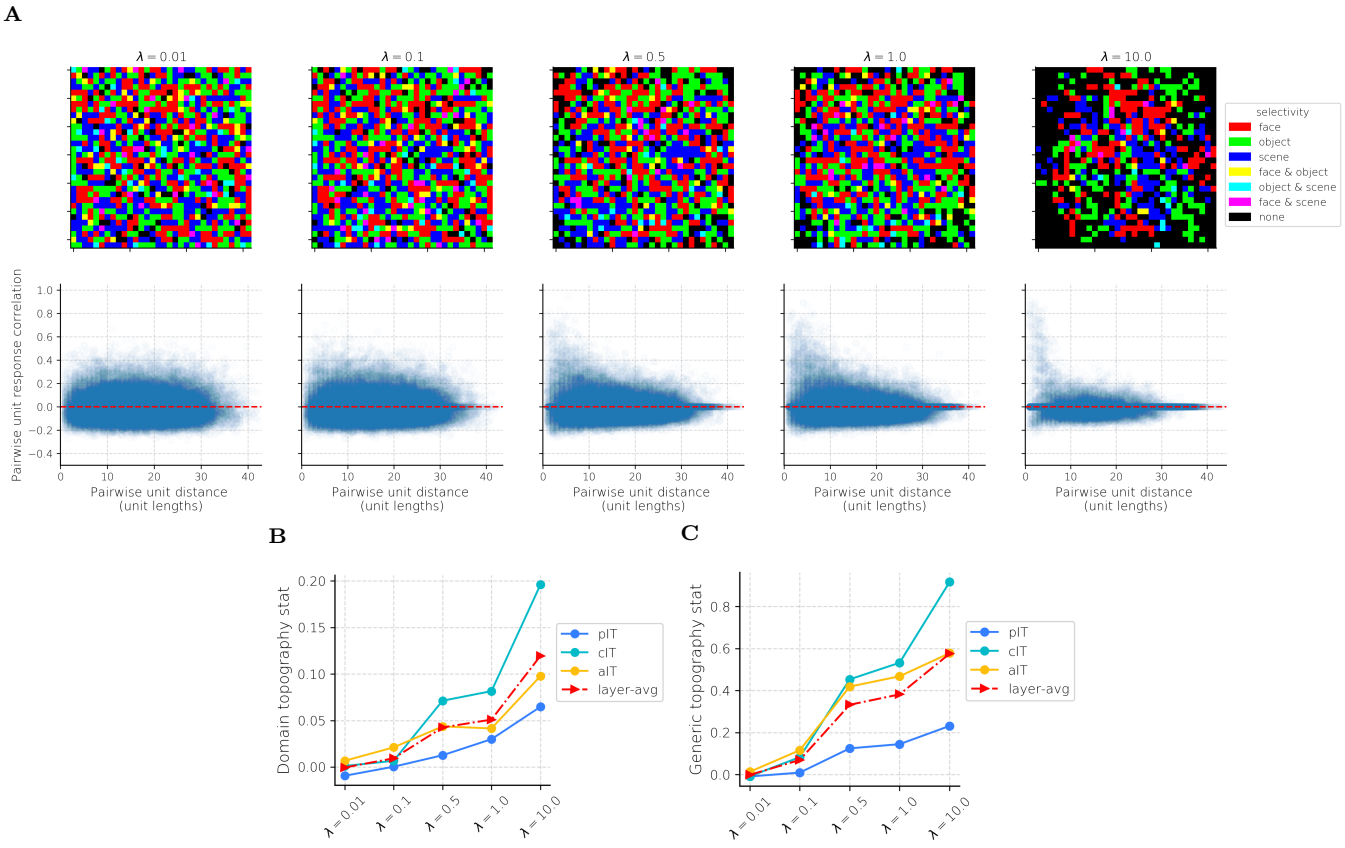


Figure S34: Domain-level and generic topography as a function of λ_w in a feedforward model without sign constraints. **A.** Emergent topography. **B.** Quantification of domain-level topography in terms of the dot product between 3-D vectors of domain selectivity, scaled by distance and summed over unit pairs.

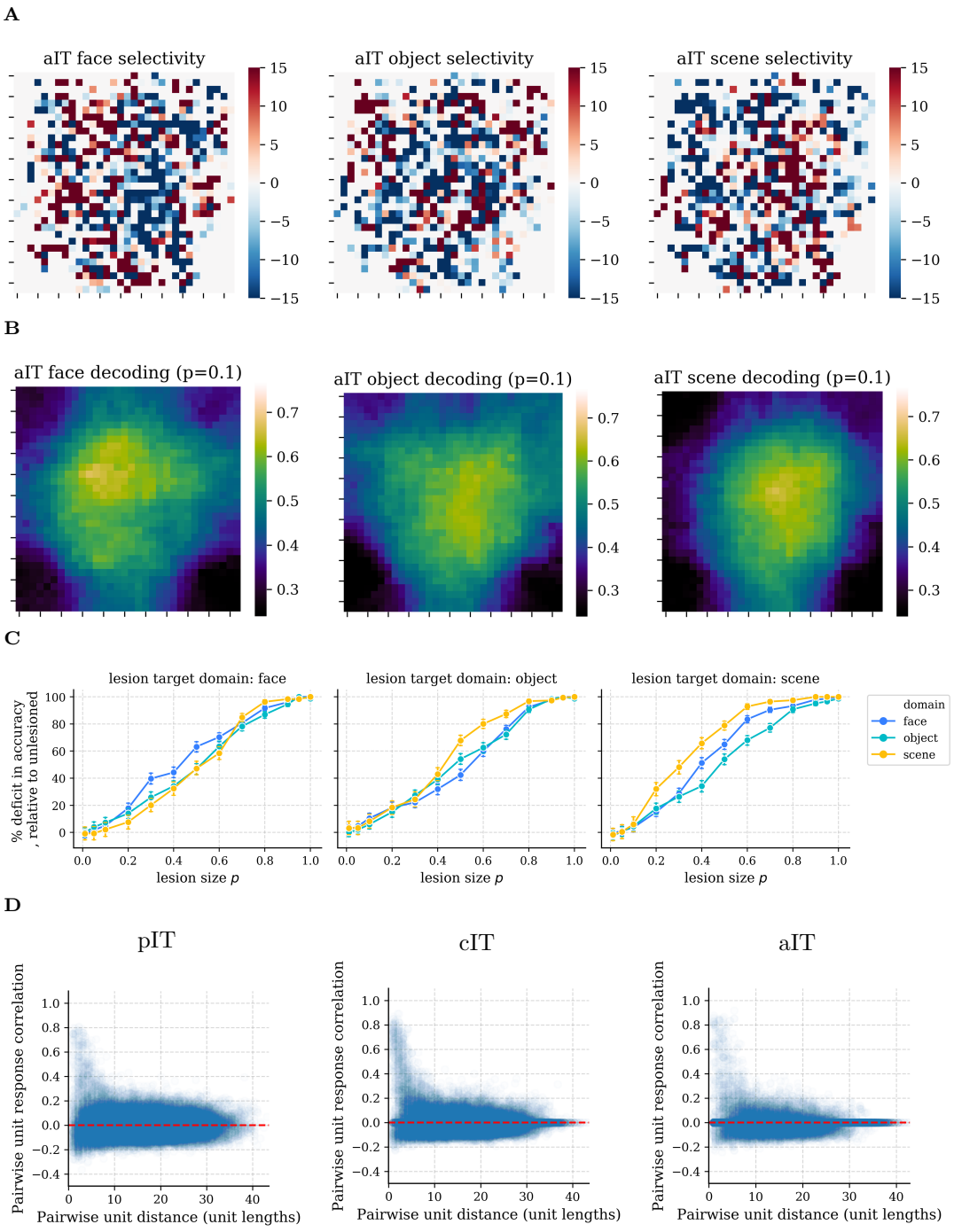


Figure S35: Results for a feedforward model without sign constraints (FNN).

20 Assessing wiring cost across architectural variants

Here, we plot wiring cost for a single wiring penalty $\lambda_w = 1.0$, to more clearly visualize how architecture impacts wiring cost. We chose to fix λ_w rather than use the optimal λ_w per architecture, as we found that λ_w had a much larger effect on wiring cost than architecture. Nonetheless, small effects of architecture can be seen when fixing λ_w, such that the models which did not produce substantial topographic organization (RNN, FNN) yielded larger wiring costs.

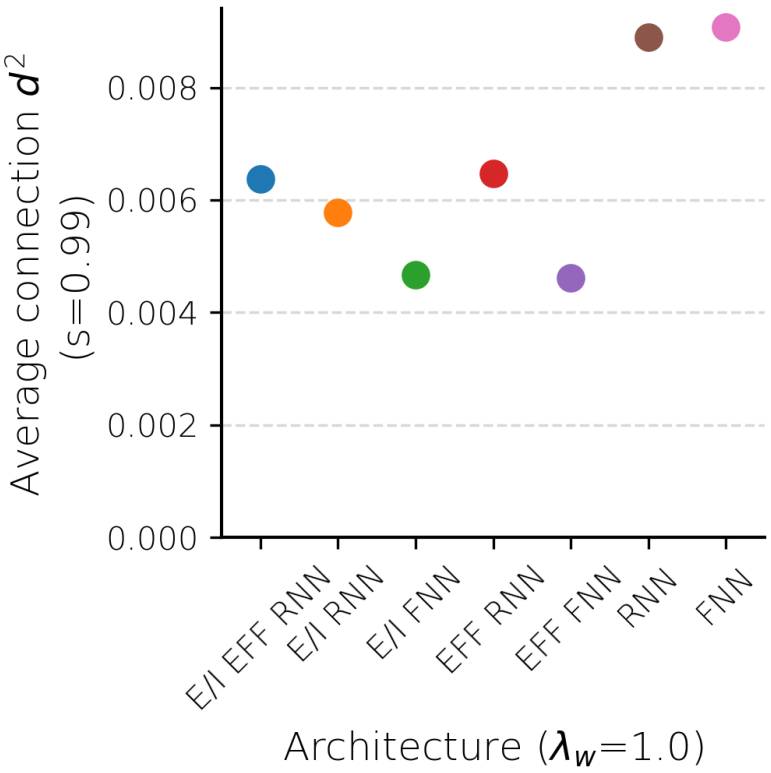


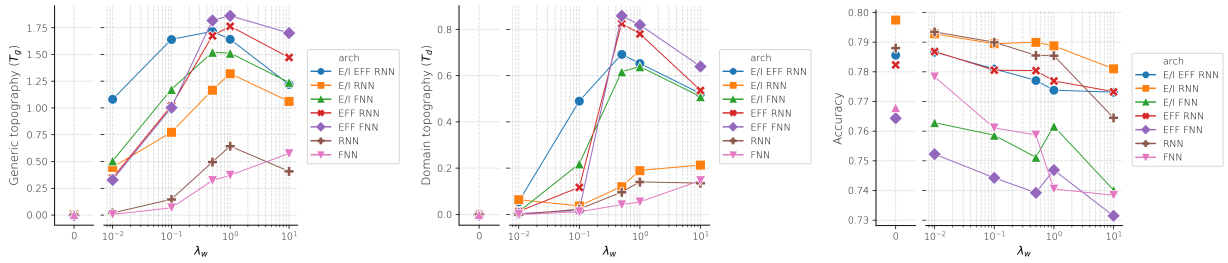
Figure S36: Wiring cost analysis for each architectural variant for fixed $\lambda_w = 1.0$, corresponding to the modal optimal λ_w value.

21 Layer normalization is important for stable training and the development of topographic organization in ITN models

Layer normalization is an important operation within the ITN models presented in the main text; it serves to standardize unit activities which makes learning much more stable, but also introduces a form of recurrent lateral processing whereby units in the same layer can influence the activity of other units in the layer. A λ_w tuning analysis for models with and without layer normalization is shown in Figure S37A and B, respectively. In contrast to the models with layer normalization, only a few models without layer normalization successfully trained to reasonable accuracy, and those that did performed worse than the corresponding models with layer normalization. However, these successfully training models (of the E/I EFF RNN and E/I RNN architectures) did produce reasonably large values of T_g and T_d.

In Figure S38 we show more in-depth results for topographic organization in an E/I RNN model with $\lambda_w = 1.0$. Intriguingly, while the model exhibits a decay in distance-dependent response correlation, with larger variance in the pairwise correlations at longer distances, indicating a larger number of units with moderate correlation at these longer distances (compare with aIT in the main model, shown in Figure 4, where the max positive correlations decay to nearly 0 at long distances). Correspondingly, while domain-level topographic clustering is seen, the domains do not organize into 3 neat areas, but rather, organize topographically with a greater spatial periodicity. This may be understood as the result of a lack of long-range (but untuned) inhibition normally introduced by the layer normalization operation. Topographic lesions indicate that faces exhibit the most topographic specialization, whereas metric-ordered lesions indicate a large but graded degree of specialization that is more graded than in the main model. Thus, layer normalization is not strictly necessary for the development of topographic organization in ITN models, but contributes strongly to training stability and global topographic organization. An interesting line of future research would be to examine more biologically-plausible methods for stabilizing activity in ITN models, possibly through modeling a diverse range of inhibitory interneuron types with different anatomical and computational characteristics.

A With layer normalization



B Without layer normalization

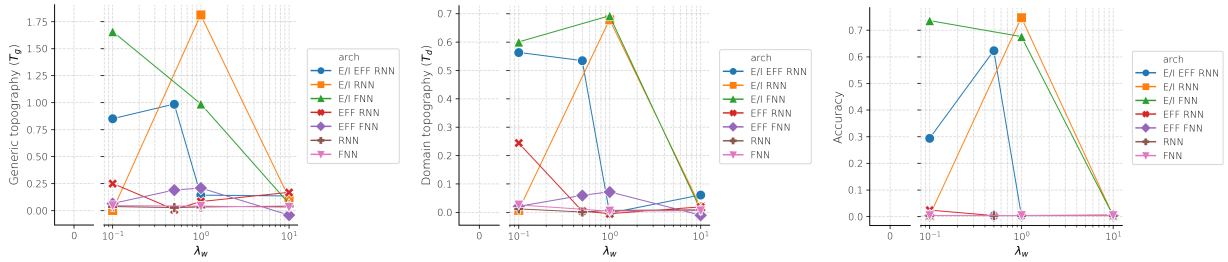


Figure S37: Topographic organization, performance, and wiring cost as a function of spatial regularization strength (λ_w) and architectural constraints, comparing across models with and without layer normalization. As in the main text, 7 architectures were tested, sweeping all unique variations of models containing or not containing: separate excitation and inhibition (E/I), excitatory-only feedforward connectivity (EFF), and learned lateral/recurrent connections (RNN vs. FNN). **A.** With layer normalization. **B.** Without layer normalization.

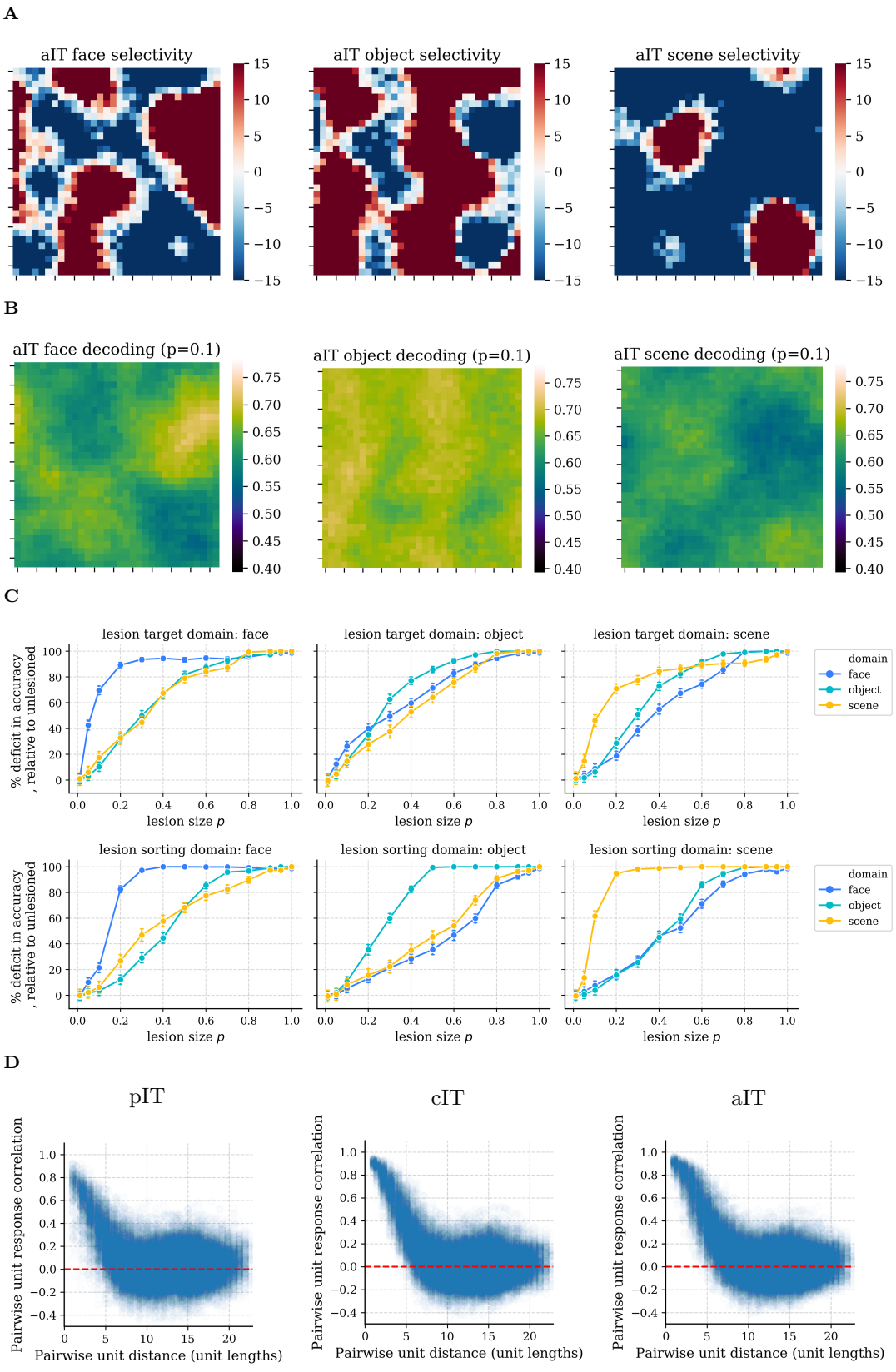


Figure S38: Removing layer normalization in a model with separation of E and I but no restriction on the sign of feedforward connectivity. Layer normalization was typically found to be necessary to get models to stably train (Figure S37), however to our surprise, we found that some instances of the E/I RNN and E/I EFF RNN models could successfully train without layer normalization. Here, we plot the subjectively best E/I RNN model using $\lambda_w = 1.0$. The results plotted in this figure demonstrate that layer normalization is not strictly necessary for topographically ordered representations in an ITN model. However, the topography appears to exhibit less domain-level global order, which may be the result of a lack of broad (but untuned) inhibition normally induced through layer normalization.

22 Connection noise aids, but is not strictly necessary for, development of topographic organization in ITN models

Similar to layer normalization, connection noise was used as a fixed mechanism across all models to improve their function. In general, connection noise makes individual neurons less reliable, and therefore forces the network to rely more on a distributed representation over all neurons, which is more affected by the spatial constraint. In this way, it is similar to Dropout and DropConnect. Here, we plot results for a version of the E/I-EFF-RNN architecture in which we removed connection noise but nevertheless examined the development of topographic organization. This model required a large wiring cost parameter $\lambda_w = 10$ and fast time constant ($a = 1.0$) to produce topographic organization (the faster time constant meant that the network had to work harder to maintain its task-relevant outputs in the maintenance period following the end of the stimulus presentation). However, like the main ITN model, it showed clustered and functionally significant domain-level selectivity and distance-dependent response correlations (Figure S39). Thus, it is more difficult—but as shown here, not impossible—to achieve topographic organization in ITN models without connection noise.

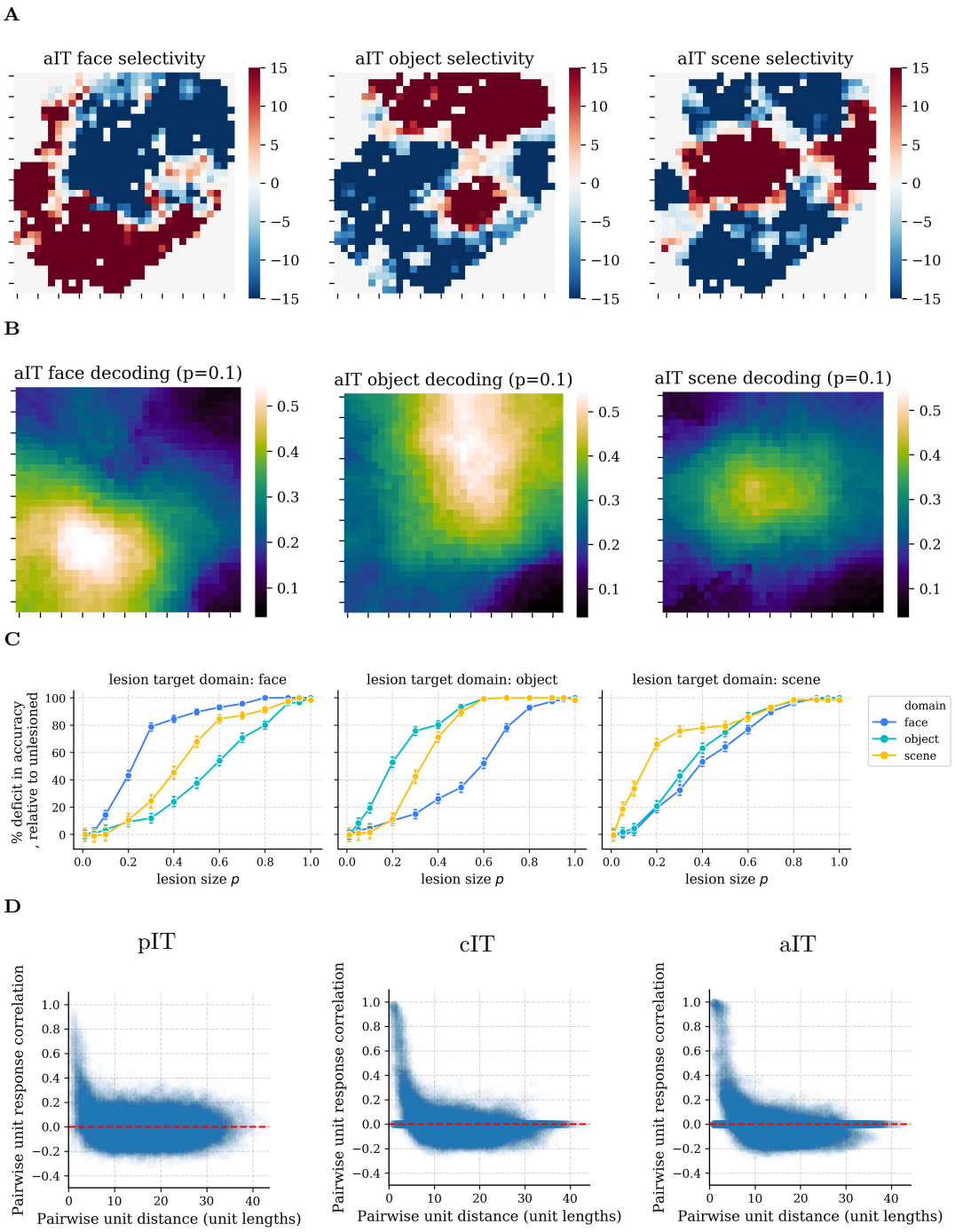


Figure S39: Removing connection noise in the main model. Topographic organization is seen, but required a larger wiring penalty ($\lambda_w = 10$) and faster time constant ($a = 1.0$).

References

Bao, P., She, L., Mcgill, M., and Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, (January 2019).

Blauch, N. M., Behrmann, M., and Plaut, D. C. (2021). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. *Cognition*, 208.

Chang, L., Egger, B., Vetter, T., and Tsao, D. Y. (2021). Explaining face representation in the primate brain using different computational models. *Current Biology*, page S0960982221005273.

Dobs, K., Martinez, J., Kell, A. J., and Kanwisher, N. (2021). Brain-like functional specialization emerges spontaneously in deep neural networks. Preprint, Neuroscience.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Konkle, T. and Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience*, 33(25):10235–10242.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2(November):4.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., and Dicarlo, J. J. (2018). CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *bioRxiv*, pages 1–9.

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L. K., and DiCarlo, J. J. (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. Preprint, Neuroscience.

Prince, J. S. and Konkle, T. (2020). Computational evidence for integrated rather than specialized feature tuning in category-selective regions. *Journal of Vision*, 20(11):1577.

Stigliani, A., Weiner, K. S., and Grill-Spector, K. (2015). Temporal Processing Capacity in High-Level Visual Cortex Is Domain Specific. *Journal of Neuroscience*, 35(36):12412–12424.

Swindale, N. V. (1982). A model for the formation of orientation columns. *Proc. R. Soc. Lond.*, B(215):211–230.

Yue, X., Robert, S., and Ungerleider, L. G. (2020). Curvature processing in human visual cortical areas. *NeuroImage*, 222:117295.