Title

Distinct roles of putative excitatory and inhibitory neurons in the macaque inferior temporal cortex in core object recognition behavior

Abbreviated title:

Probing the role of IT cell types in object recognition

Authors

Sachi Sanghavi[1,2], Kohitij Kar [3,4,5,6*]

Affiliation

1. Visual Intelligence and Technological Advances Laboratory, York University, Toronto, Ontario, Canada
2. Department of Computer Sciences, University of Wisconsin-Madison, USA
3. Department of Biology, Faculty of Science, York University, Toronto, Ontario, Canada
4. McGovern Institute for Brain Research at Massachusetts Institute of Technology, Cambridge, MA, USA
5. Vision: Science to Applications (VISTA), York University, Toronto, Ontario, Canada
6. Centre for Vision Research, York University, Toronto, Ontario, Canada

[*]Correspondence should be addressed to Kohitij Kar.
E-mail: k0h1t1j@yorku.ca

Conflict of interests

The author declares no competing financial interests.

Acknowledgments

KK was supported by the Canada Research Chair Program and the Simons Foundation Autism Research Initiative (SFARI)-Pilot award. This research was undertaken thanks in part to funding from the Canada First Research Excellence Fund. We also thank Jim DiCarlo for his support.

Abstract

A spatially distributed population of neurons in the macaque inferior temporal (IT) cortex supports object recognition behavior, but the cell-type specificity of the population in forming "behaviorally sufficient" object decodes remain unclear. To address this, we recorded neural signals from the macaque IT cortex and compared the object identity information and the alignment of decoding strategies derived from putative inhibitory (Inh) and excitatory (Exc) neurons to the monkeys' behavior. We observed that while Inh neurons represented significant category information, decoding strategies based on Exc neural population activity outperformed those from Inh neurons in overall accuracy and their image-level match to the monkeys' behavioral reports. Interestingly, both Exc and Inh responses explained a fraction of unique variance of the monkeys' behavior, demonstrating a distinct role of the two cell types in generating object identity solutions for a downstream readout. We observed that current artificial neural network (ANN) models of primate ventral stream, designed with AI goals of performance optimization on image categorization, better predict Exc neurons (and its contribution to object recognition behavior) than Inh neurons. Beyond, the refinement of linking propositions between IT and object recognition behavior, our results guide the development of more biologically constrained brain models by offering novel cell-type specific neural benchmarks.

Keywords

Inferior temporal cortex, cell types, population decoding, deep neural networks, object recognition

Introduction

Distributed neural population activity in the macaque inferior temporal (IT) cortex, which lies at the apex of the visual ventral stream hierarchy, is critical in supporting an array of object recognition behavior[1–3] . Previous research, however, has been largely agnostic to the relevance of specific cell types, inhibitory (Inh) vs. excitatory (Exc), in the formation of "behaviorally sufficient" IT population codes[2] that can accurately predict primates' image-by-image object discrimination patterns. In this study, we aim to identify how these distinct classes of neurons in the IT cortex jointly contribute to object recognition behavior.

In the mammalian neocortex, approximately 15-25% of neurons are GABAergic inhibitory neurons, with the majority being glutamatergic excitatory neurons[4–6]. A rich body of research suggests that inhibitory and excitatory neurons play distinct roles in shaping the functional representation of images in the IT cortex. While some studies have revealed a comparable level of stimulus selectivity between these two classes of neurons[7], others propose a higher degree of stimulus selectivity in excitatory neurons[8]. In addition, local inhibitory circuits are also shown to shape the stimulus selectivity of excitatory neurons in TE[9]. We argue that specific stimuli choices, brain areas, and methods of neuronal recordings can limit the scope of such neuronal characterization without explicitly linking these properties with a specific behavior. Instead, we propose leveraging the behavioral relevance of cell types, across a battery of tasks, to reconcile these diverse findings and motivate an approach to ground the studies of their functional roles (the term "function" relating to relevance in a behavior rather than an interpretable stimulus or neuronal property).

A falsifiable way to generate insights on how diverse neuronal populations in the IT cortex represent visual stimuli and participate in specific behaviors is to develop image-computable artificial neural network (ANN) models of the ventral stream and the behaviors it supports[10]. While ANN models from the family of deep convolutional neural networks explain primate object recognition behavior at unprecedented levels, these models currently present many shortcomings[11–13]. What are the visual system's most functionally relevant properties lacking in these models? Of note, current ANN models of the ventral stream[14,15] do not explicitly represent neuronal cell-type. This is in contrast with other attributes of the IT cortex like connectivity patterns[16,17], functional topography[18–20], learning mechanisms[21,22], and hypothesized evolutionary training objectives[23,24] that has been recently considered to augment models of object recognition. Given the distinct nature of the Exc/Inh ratio in the cortex[4,5], and the previous reports of differences in their responses to visual stimuli[7,8], we reasoned that understanding the cell-type specific constraints will be critical to developing more brain-aligned models of object recognition behavior.

In this study, we performed large-scale neural recordings while monkeys fixated images presented in their central field of view. Monkeys also performed 28 different binary object discrimination tasks. We identified putative inhibitory and excitatory neurons in the IT cortex using spike shape-based sorting techniques. We then compared the strength of behavioral predictions of neural decoding ("readout") models constructed from specific (putative) cell types in the IT cortex. Overall, we observed that decoding strategies derived from excitatory neurons significantly outperform those produced by inhibitory neurons in overall accuracy and image-by-image match to monkey behavioral patterns. Interestingly, we observed that inhibitory and excitatory neurons

explained partially overlapping as well as unique fractions of the monkeys' behavioral variance, emphasizing their distinct roles in object recognition behavior. While evaluating how current ANN explain the response patterns of the specific IT cell types, we observed that all tested ANNs predicted Exc neurons significantly better than Inh neurons. In sum, our results provide correlative evidence for cell-type specificity in the linkage between IT population activity and object recognition behavior, along with the novel cell-type specific constraints for guiding the next generation of brain models.

Results

As outlined above, we first recorded visually driven neural activity in the macaque IT cortex and categorized them into two separate classes based on their spike waveform shape: narrow spiking (putative inhibitory) and broad spiking (putative excitatory) neurons. Below, we first report the diversity in the response properties of these neuronal cell types and then link these responses to the measured behavioral variances in the macaques. Finally, we evaluate how current ANN models of ventral stream explain these neural responses.

Identification of putative excitatory and inhibitory neurons

Multiple neurophysiological studies have shown that inhibitory neurons elicit short-duration action potentials[25–28] which are discharged at higher frequencies. This has led to the term fast spike (FS)[29] or "thin spikes"[30] for inhibitory neurons, compared to regular spike[30] (RS) for excitatory neurons. Though a small portion of inhibitory neurons exhibit broader action potentials[27,31], and a small fraction of excitatory neurons have narrower action potentials[32,33], this physiological difference of narrow vs. broad spikes has been widely used to classify extracellularly recorded neurons in several primate neocortical areas, including V4 [34], inferior temporal (IT) cortex [8,35], and prefrontal cortex (PFC) [36–39]. Therefore, we leveraged similar spike shape-based criteria (see Methods for details) to classify neurons into putative excitatory and inhibitory cell types. In brief, as shown in **Figure 1A**, we recorded neural activity across the macaque inferior temporal (IT) cortex while they passively fixated on images presented for 100 ms each. After the filtered data was thresholded to detect the multiunit spikes, we performed principal component analysis (PCA) to cluster the spikes into individual neurons per recording site (**Figure 1B**). Then we classified these neurons as narrow and broad spiking neurons based on their spike waveform shapes (specifically using the trough to peak and repolarization time as shown in **Figure 1C**). Our analysis yielded 124 broad spiking (putative excitatory, Exc) and 37 narrow spiking (putative inhibitory, Inh) neurons (**Figure 1D**). Interestingly, as these numbers demonstrate, ~23% of our pooled neurons across monkeys could be classified as putatively inhibitory, which is consistent with prior reports[5,6].

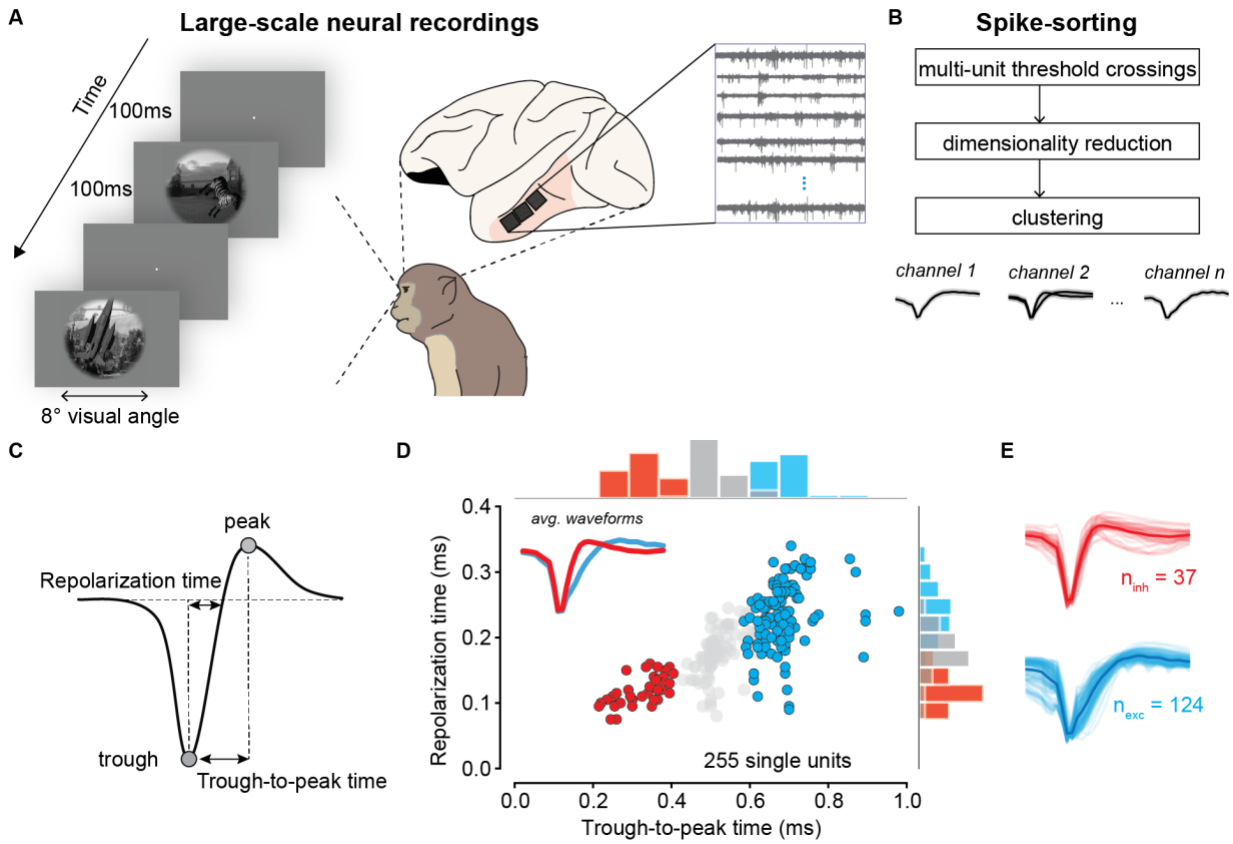

Figure 1. Neural recordings and classification of neurons. (**A**) We measured neural responses from inferior temporal (IT) cortex using chronically implanted multi-electrode Utah arrays while the monkeys passively fixated on images (n=640). Inset shows a schematic of bandpass filtered voltages across neural recording sites. (**B**) Spike-sorting steps to convert neural responses from multi-unit sites into single units. (**C**) Illustration of the two action potential (AP) features (trough to peak time, and repolarization time) used for classifying neurons into broad spiking (putative excitatory, Exc) and narrow spiking (putative inhibitory, Inh) types. (**D**) 2D scatter plot showing k-means clusters of all identified single-units (n=255) on the two AP features in (C). Putative inhibitory neurons (n=37) are in red, putative excitatory neurons (n=124) in blue. Gray points are excluded from all further analyses.

Putative excitatory and inhibitory neurons show distinct functional response properties

To ensure that the experimental noise did not affect the recordings of the Exc and Inh neurons differentially, we first compared the image rank-order response reliabilities (see Methods) of the Inh and Exc neural responses. We observed no significant differences (ΔReliability = 0.02, t-test: t(159)=1.1159, p = 0.2661) between these two populations, and a relatively high SNR (Reliabilities = 0.82 and 0.8 for Inh and Exc respectively) across these two neural populations. To test whether the putative Exc and Inh neurons demonstrate known physiological properties we first compared the baseline and maximum evoked firing rates of these two neural populations. Consistent with previous research[8,26], we observed that both baseline (t-test, Exc < Inh, t(159)=-2.425, p<0.001)

and evoked (t-test, Exc< Inh, t(159) =-3.72, p<0.001) firing rates were significantly higher for the Inh neural responses (**Figure 2A**; Exc: Blue, Inh: Red). We also observed that the overall response variance is significantly higher (Exc< Inh, t-test, t(159)= -4.60, p<0.001) in Inh neurons (normalized histogram shown in **Figure 2B** for both Exc: blue and Inh: red neurons). This finding is similar in nature to broader tuning preferences for stimulus orientation and spatial frequency observed in inhibitory neurons in the mouse [40] and cat[41] visual cortex. In contrast, Exc neurons exhibited a significantly higher level of category selectivity (t-test, t(204) = 17.53, p<0.001 ; see Methods for definition). **Figure 2C** shows a histogram of difference between category selectivity (Exc – Inh) across different time bin of responses (see **Figure 3A**). This result demonstrates that Exc neurons tend to respond more similarly to images from the same object category compared to other objects. Therefore, it is likely that such representations might produce more linearly separable object representations.

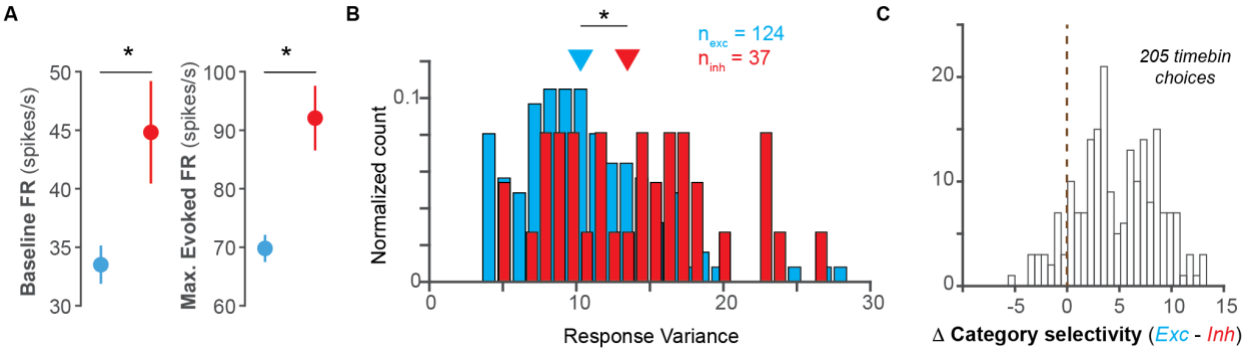

Figure 2. Functional response properties of Exc and Inh neurons. (**A**) Baseline and max evoked firing rates (mean ± SD across neurons) for Exc (blue) and Inh (red) neurons. Both firing rates were measured between 70 and 170ms from stimulus onset (see Methods). Unpaired two-tailed two-sample Welch's *t*-test show higher baseline and evoked firing rates for Inh neurons. (**B**) Histogram of response variances. Response variance, computed per neuron, is the variance in mean response in 70-170ms interval across all 640 images. (**C**) Histogram of difference in category selectivity between Exc and Inh neurons.

Putative excitatory neurons produce higher accuracy in object discrimination

Given that we observed that Exc neurons exhibit higher object category selectivity, we next tested whether constructing linear object identity decoders based on Exc (ticks and population vector in blue; **Figure 3A**) could lead to higher accuracies on an object discrimination task compared to Inh neurons (H_1**, Figure 3B**). The other possibilities are that Inh neuron-based population decodes outperform the Exc based ones (H_2**, Figure 3B**), and that these two populations produce equally accurate decoders (H_0**, Figure 3B**). In these analyses, the temporal integration window for spike counts is typically a free parameter determined by the start time of the window (tStart) and the binwidth size (**Figure 3A**). We constructed 205 decoding models (corresponding to different population vectors) based on varied choices of tStart and binwidth. We observed that the decoders that produced the higher accuracies were constructed from the Exc-based population vectors (supporting H_1). The strongest decoders were constructed by integrating Exc neural responses from 90 to 190 ms from Test image onset (**Figure 3C**). The accuracies when measured in this fashion

depend heavily on the number of neurons sampled. Therefore, we sampled from 2 to 30 neurons (randomly for both Exc and Inh pools). The difference in accuracies between the two groups increased as a function of the number of neurons (**Figure 3D**). Therefore, we expect our reported number as a lower bound. While Exc outperformed Inh based decodes, it is important to note that the linearly separable category information contained in the representation of the Inh neurons is significantly greater than zero (with a d' ~2) and only 9.31% (mean) lower than the Exc neurons. Therefore, we further probed how the decoding accuracies evolve as we vary the ratio of participation of the Exc and Inh neurons in the neural population pool (keeping the overall number of neurons constant at n=30). Our results (**Figure 3E**) show that decoding accuracies increase as a function of the Exc/Inh ratio with the inclusion of more Exc neurons.

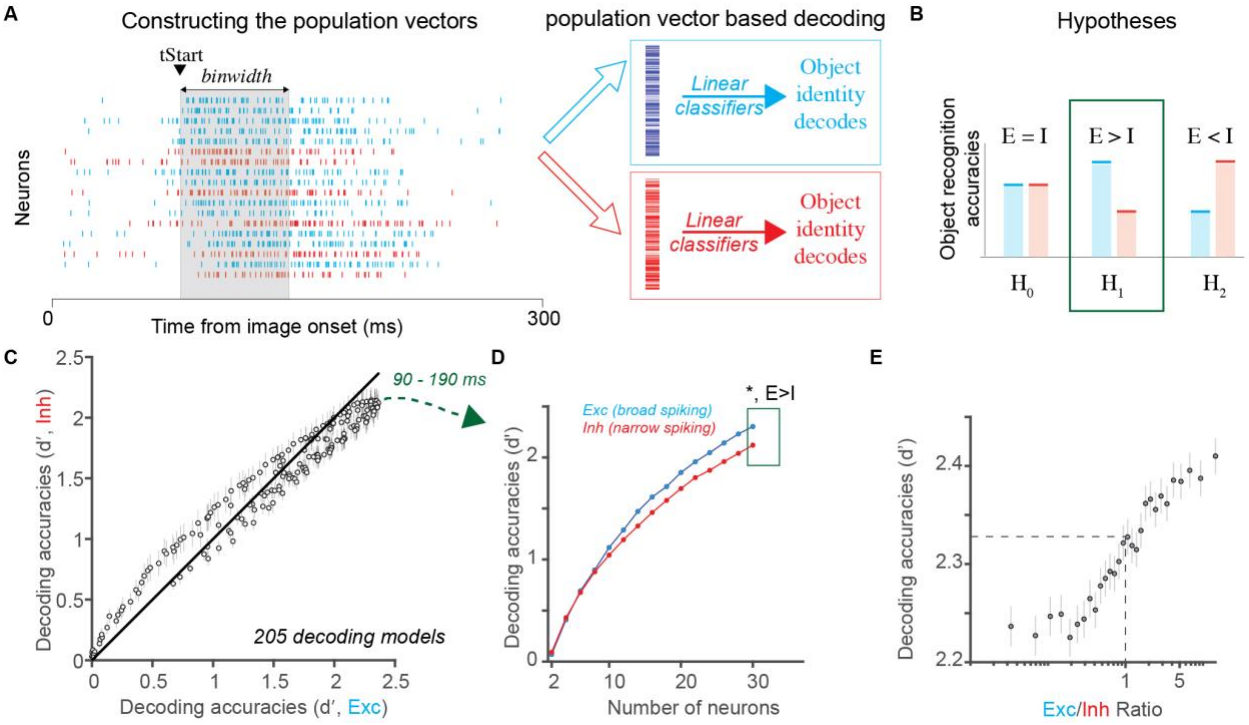

Figure 3. Performance of Exc and Inh population based decoders on object recognition. (A) We trained 205 linear decoders (with varying tStart and binwidth) on Exc and Inh populations for object recognition test. Each decoder was trained on the mean response of neurons in a population. Since the mean response is determined by the number of spikes in some interval characterized by two parameters, namely the start time with respect to stimulus onset (tStart) and the bin width (binwidth), we constructed multiple decoders by varying this interval. (**B**) One hypothesis (H_0) is that linear decodes from both populations would do equally well on object identity discrimination. Another hypothesis (H_1) is that Exc based decodes would perform better than Inh based decodes; or alternatively (H_2), Inh based decodes would perform better than Exc based decodes. (**C**) Performance of decoders trained on Exc and Inh populations. Each data point is a linear decoder (n=205) constructed by varying tStart and binwidth. Error bars represent SD over random sampling of neural population (n=30). (**D**) Performance of the 'best' decoder (90-190 ms) as we varied the total number of (Exc vs. Inh) neurons sampled. (**E**) Performance of the 'best' decoder as we varied the Exc/Inh ratio. Total neurons was always kept at 30. Error bars represent SD over random sampling of neural population.

Putative excitatory neuron based decodes produce higher image-level consistency with monkeys' object discrimination behavior.

While the decoding accuracies estimated above allow us to quantify the overall object information (that is linearly separable) represented in the Exc and Inh neural populations, it doesn't necessarily render them behaviorally relevant. To test the relevance of the cell type specific neural decodes towards the monkeys' behavioral reports, we first measured the behavioral responses in 3 macaques on 28 delayed match to sample object discrimination tasks. On each trial, monkeys viewed a Test image (containing a target object; one among 8), followed by a choice screen containing canonical views of two objects (the target and a distractor object). The monkey had to indicate their choice of the target object with a saccade to one of the options (**Figure 4A**). We then estimated an image-level metric (referred to as I_1) from the measured behavioral performances. In brief, the I_1 metric reflects the performance of the monkey in identifying the target object in each image, when tested across all possible distractor categories (see Methods, also for more detail see the estimation of B.I_1 in[11]). The trial level split-half reliability of this metric, after behavioral trials were pooled across 3 monkeys (yielding 204 repetitions per image) was 0.89 ± 0.004 (mean ± SD). We quantify the consistency of the neural population based predictions (N.I_1) with the behaviorally estimated I_1 as the trial-level noise corrected correlation between the two. Similar to the overall accuracies, we also observed that Exc neuron based decodes (**Figure 4C**) yielded significantly higher behavioral consistencies (paired t-test; $t(204) = 47.63$, $p<0.0001$) compared to Inh neurons (supporting H_2, **Figure 4B**).

Given that the inhibitory neurons represent a significant fraction of object information and that these decodes correlate highly with our measured behavioral metrics, we further invested whether combining Exc and Inh neurons in the population vector produce more behaviorally aligned decodes. In sharp contrast to the decoding accuracies (**Figure 3E**), where adding Exc neurons always increased decoding accuracies, the consistency results show that the benefits of adding Exc neurons to the population code asymptote around a Exc/Inh ratio of ~3-4 (**Figure 4D**). This suggest that both populations might represent a unique fraction of the behavioral variance and that including Inh neurons to the population code might be necessary for fully explaining the behavior. Next, we test this hypothesis more directly.

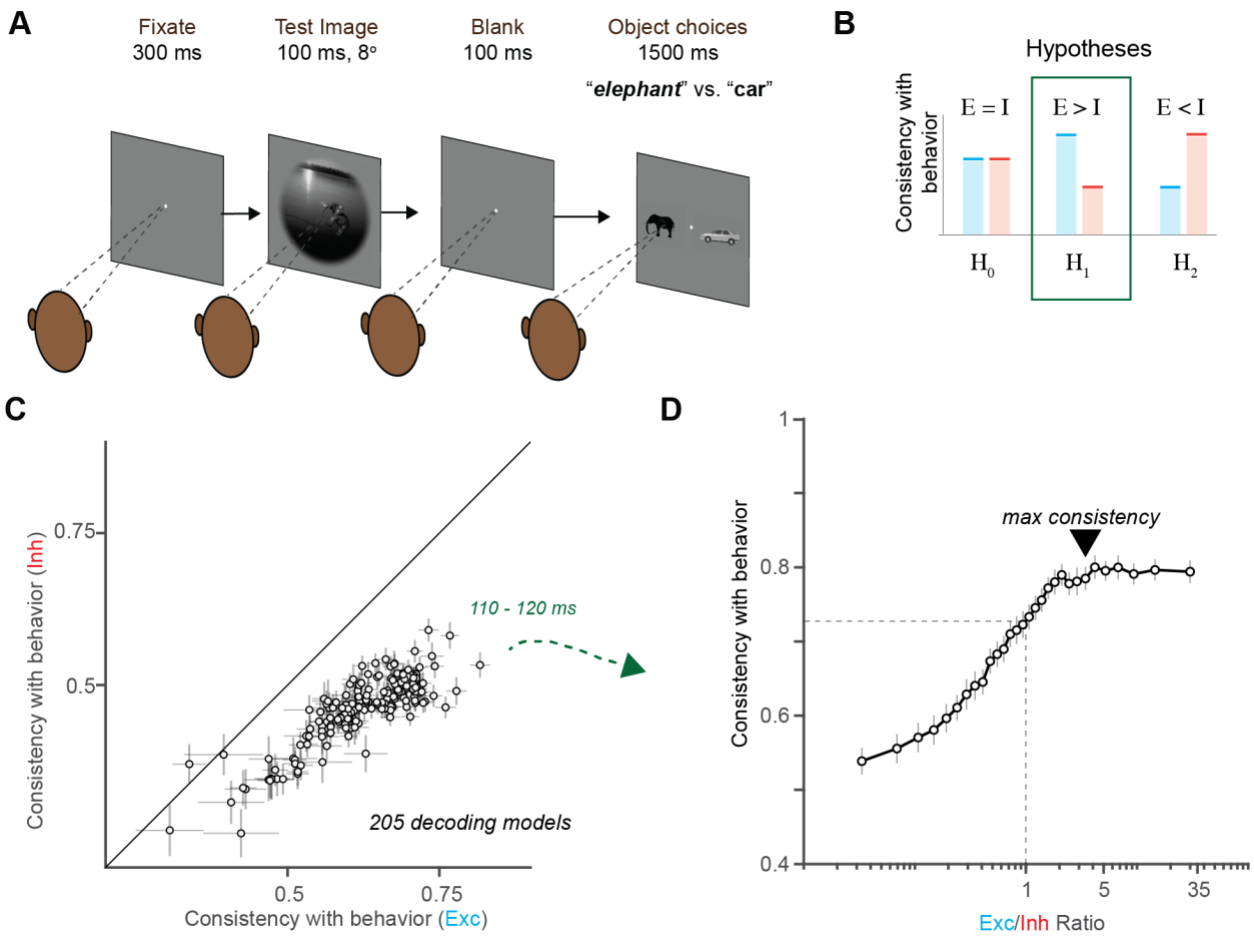

Figure 4. Comparing monkey behavior with the IT population (Exc vs. Inh) based readout predictions. (A) We tested 3 monkeys on multiple delayed match to sample binary object discrimination tasks. Monkeys initiated a trial by fixating on a central white dot for 300 ms, following which a test image (out of 640 images) containing the target object (shown elephant) was presented in the central 8°. This was followed by a blank screen for 100 ms, and then canonical views of the target (elephant) and distractor (car; one out of 7) object were shown on the screen (randomly assigned distractor for each trial). The monkeys had 1500ms to freely view the object choices and indicate their selection of the target by fixating on it for 400 ms. (**B**) One hypothesis (H_0) is that linear decodes from both populations would be equally consistent with monkey behavior. Another hypothesis (H_1) is that Exc based decodes would be more consistent with monkey behavior; or alternatively (H_2), Inh based decodes would be more consistent with monkey behavior. The green rectangle indicates the hypotheses that the data couldn't falsify. (**C**) Behavioral consistency of Exc and Inh based decodes. Each data point is a linear decoder (n=205) based on a specific choice of tStart and binwidth (refer Figure 3A), and error bars denote SD across random draws of 30 neurons from the Exc and Inh sets respectively. The highest consistency readout was from Exc neural response in 110 to 120 ms (relative to Test image onset) interval. (D) Behavioral consistency of the highest consistency decoder found in (C) for varying Exc/Inh ratio (with a fixed number of neurons. Max behavioral consistency is achieved for E/I ratio ~3-4, after which it asymptotes. Error bars denote random sampling of neurons to construct the neural pool (size=30). Horizontal dashed line indicates the ceiling, that the internal consistency of the monkeys' behavioral I_1 (mean ± SD).

Putative excitatory and inhibitory neurons explain non-overlapping variances in object discrimination behavior

To test how much behavioral variance is uniquely explained by the Inh and Exc populations, we performed a partial correlation analysis. **Figure 5A** shows a comparison of the Exc based decodes with (lighter blue circles) and without (darker blue circles) controlling for the inhibitory neural decodes for the same time bin choices (205 choices produced by varying tStart and binwidth; refer **Figure 3A**). We observed that, indeed a significant fraction of behavioral variance (max ~ 12%) explained by Exc neuron based decodes was reduced when controlled for Inh neurons. Similarly, Inh based decodes also showed an exclusive fraction of variance preserved after the Exc contributions were controlled for (**Figure 5B**). We observed that the Exc cell-type exclusive behavioral variances peaked between 90-130 ms (inset; **Figure 5C**). Our results suggest that readouts for behavior from the IT population likely engage a temporally integrative mechanism that optimally combines information from these two populations. Therefore, we argue that brain models that do not adequately explain the unique variances represented by the Exc and Inh neurons will be insufficient to explain primate behavioral error patterns.

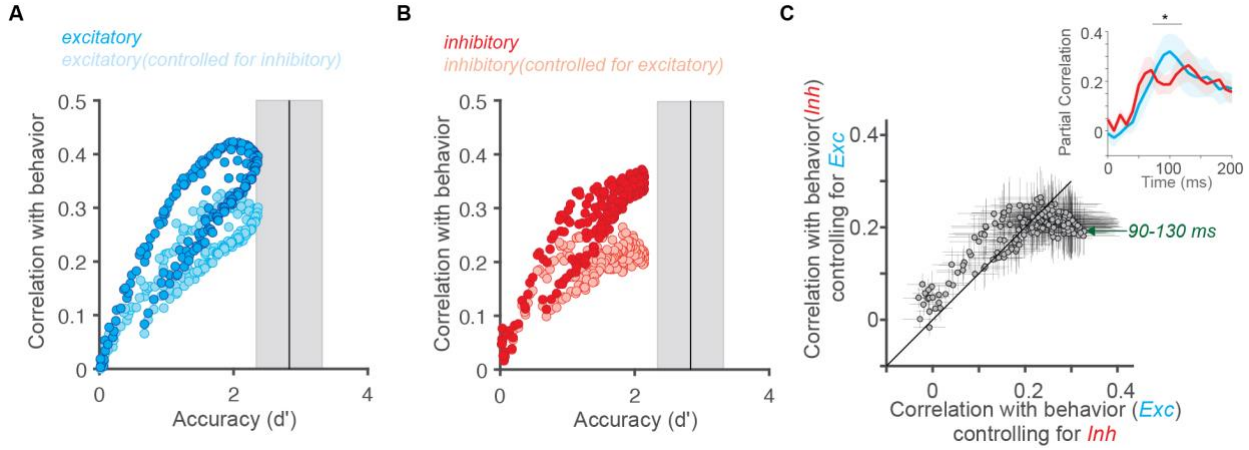

Figure 5. Partial correlation analysis between monkey behavior and readouts from Exc and Inh neurons. (**A**) Correlating monkey behavior and Exc based decodes ($N.I_1$, n=205) with (lighter blue circles) and without (darker blue circles) controlling for readouts from Inh neurons. Vertical black line represents average monkey performance, and shaded region is SD. (**B**) Correlating monkey behavior I_1, and Inh neurons based decodes ($N.I_1$, n=205) with (lighter red circles) and without (darker red circles) controlling for readouts from Exc neurons. (**C**) Comparing correlation of monkey behavior with cell-type unique variances (Exc vs. Inh). Error bars represent SD over random sampling of neural population (n=30). Inset shows the time of the unique variances. Time bin considered was 20 ms.

ANN models of object recognition better predict putative excitatory than inhibitory neurons

Our results suggest that putative excitatory and inhibitory neurons explain a unique fraction of the behavioral variance across time (**Figure 5**). Therefore, next we asked whether current ANN

models of the ventral stream explain Exc and Inh neurons with the same prediction accuracy. To test this, we first tested a state-of-the-art brain-aligned model of the ventral stream, CORnet-S[15] that has a fixed, committed mapping between macaque ventral stream areas (V1, V2, V4, and IT) and model layers. We then used linear regression based standard measures of mapping the features of the ANN-IT layers onto the responses of individual macaque-IT neurons[16,22]. We took the goodness of fit to be the median predictivity across all recorded neural sites. Interestingly, we observed that the CORnet-S IT layer predicted Exc neurons significantly better than Inh neurons (**Figure 6A**, t-test, t(159)==2.051, p=0.04). We also observed similar trends across all tested layers of CORnet-S (dashed bars). To further test the generalizability of this result across other ANN models, we sampled 19 high-performing representative ANNs (e.g., AlexNet[42], CORnet-S[15], ResNets[43] etc.; see Methods for all model details, **Figure S1**) from Brain-Score[14]. Using Brain-Score's public dataset, we first established a fixed mapping between a specific model layer (see **Figure S1**) and the macaque IT. Consistent with results from CORnet-S, we observed that the ANN-IT layer of all selected models predicted Exc neurons better than Inh neurons (**Figure 6B**; paired t-test, t (18) = 24.0477, p<0.001). Does this gap in neural predictivity lead to a commensurate gap in the ANN's ability to predict the unique behavioral variances accounted for by the two populations?

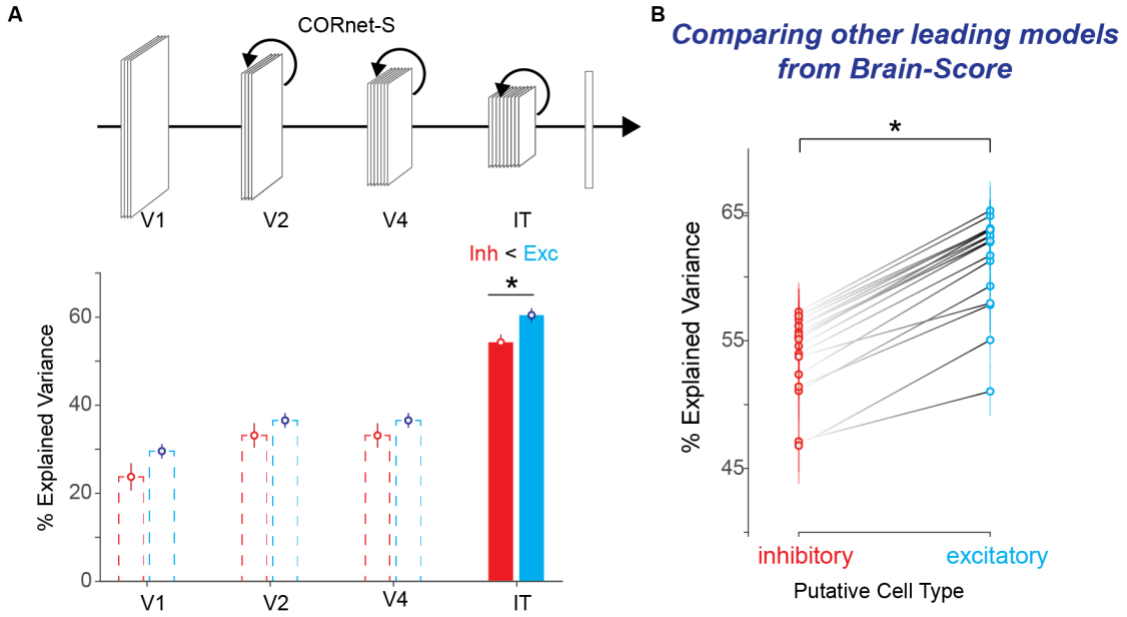

Figure 6. IT Predictivity of ANN models for Exc and Inh neurons. (A) The IT layer of CORnet-S, a model designed to match the ventral visual stream hierarchy (V1, V2, V4, and IT), predicts Exc IT neurons significantly higher than the Inh neurons (paired *t*-test; *: p < 0.05). (**B**) Comparison on IT predictivity (Exc vs. Inh) across 19 ANN models from Brain-Score. All models predicted Exc neural responses higher than Inh neurons. Error bars denote s.e.m. across neurons

Figure 7A demonstrates the intuition behind a variance partitioning analysis to address this question. The gray shaded region in the left panel of **Figure 7A** refers to the behavioral variance explained by the ANNs. In our study such estimates are performed by a correlation between ANN-based I_1 predictions and the measured behavioral I_1 vector. Next the green region in the mid panel

refers to the behavioral variance explained by the ANNs controlling for a neural decode based variance (e.g., from Exc or Inh neural population decodes). Such an estimate is retrieved by performing a partial correlation analysis between ANN I_1 and Behavior I_1 with Neural I_1 as a control variable. Finally, the right panel is the result of subtracting the first two estimate and provides the image-level behavioral variance explained by ANNs that is common with the neurally explained variance. **Figure 7B** shows the result of this analysis (across 53 models). We observe that indeed, consistently across all models, the Exc predicted variance is better captured by ANNs compared to Inh predicted variance.

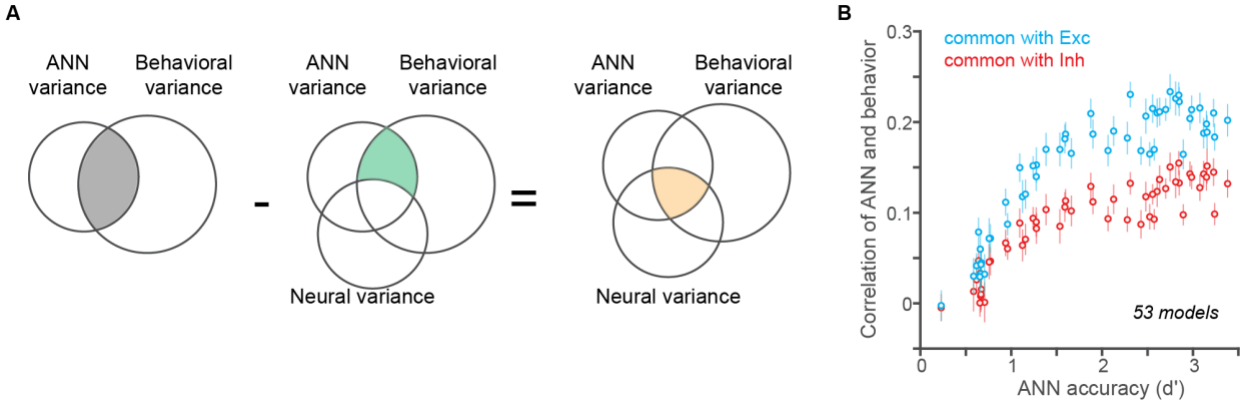

Figure 7. Variance partitioning analysis to estimate amount of behavioral variance jointly explained by neurons (Exc vs. Inh) and ANN-IT layers. (A) Venn diagrams showing shared variance. Left panel: The gray shaded region refers to the behavioral variance explained by the ANNs. Mid panel: The green region refers to the behavioral variance explained by the ANNs controlling for a neural decode based variance. Right panel: The yellow region refers to the result of subtracting the first two estimates and provides the image-level behavioral variance explained by ANNs that is common with the neurally explained variance. (**B)** Comparison of correlation of ANN predictions with behavior that is common with predictions produced by Exc (blue) and Inh (red) neurons (across 53 models).

Discussions

The primary goal of this study was to probe the cell type specificity in how the population activity of the macaque inferior temporal cortex supports core object recognition behavior. Our results provide evidence towards a unique contribution of both putative inhibitory and excitatory neurons in shaping the monkeys' behavioral reports. Below we discuss our results with respect to its broader implications.

Exc/Inh contributions to linking propositions and downstream readout mechanisms

We demonstrate that Exc and Inh neurons provide unique contributions to the object identity solutions. It is still unclear how these distinct sources of information is readout and conveyed to the downstream brain areas like the PFC[44], caudate[45] and amygdala[46]. The exact nature of those

readouts and therefore the individual contributions of the Exc and Inh neurons might be specific to the behaviors under study (e.g., object detection, facial emotion discrimination, decision-making: go no-go tasks). Therefore, our study provides a template for generating linking hypotheses to test how sensory representation across cell types relate to behavioral tasks. Given the temporal variation in the amount of object information in the Exc and Inh neurons (inset **Figure 5C**), we speculate that a dynamic integration based decoding scheme might be necessary to fully explain the object recognition behavior.

Incorporating the role of Exc/Inh contributions in ANN-based brain models

While it has remained commonplace to seek an intuitive understanding of what image based features are differentially processed by these individual cell types, such inferences are often limited by the low-dimensional, apriori hypotheses (like size of receptive fields[47], stimulus parameter tuning[9]) constructed by the experimenters. Complimentary to this approach, our data provides two types of image-level constraints that might more effectively guide the next generation of brain models. These constraints are 1) reliably estimated image by image neural response of individual Exc and Inh neurons to a wide range of images 2) image level differences in population decodes constructed from Exc and Inh neural activity. Access to these additional, finer grained, unbiased measurements could serve as target objectives in finetuning the ANN neural representations. A recent promising line of work[48] has incorporated Dale's principle (i.e., presynaptic neurons can only project as excitatory or inhibitory inputs on their postsynaptic partners) in a standard ANN architecture. Our results reported here serve both as objective benchmarks for evaluating such models. But as mentioned earlier, this data can also be used during model development to regularize the networks, as demonstrated in previous work[49].

Implications for theories of Exc/Inh imbalance in various neurological disorders

A change in the Exc/Inh ratio is often deemed as an important neural marker for many neurological disorder [50,51]. Similarly atypical sensory processing has also been a hallmark of many neurological disorders like autism spectrum disorder[52], ADHD[53], etc. Taken together, we speculate that understanding how participation of the Exc and Inh neurons shape the behavior of neurotypical adults, might be key in developing more robust theories of the differences that lead to clinical symptoms [54]. Selective targeting of Exc or Inh neurons with genetic techniques might help to model these disorders in experimental settings, uncovering novel therapeutic strategies. In this regard, it is important to also expand our study design to incorporate more specific cell types[55] beyond the broad classifications of Inh and Exc neurons.

Future directions: Selective perturbation of excitatory and inhibitory neurons to test circuit level hypotheses

Our results allow us to generate testable predictions for future direct brain perturbation experiments that could selectively modulate Exc and Inh neurons. Recent advances in optogenetic[56] and chemogenetic[57] strategies allow for selective transfect excitatory and inhibitory neurons and alter their responses. Such methods will allow us to explicitly test whether selective perturbation of Exc or Inh neurons can produce network changes like the balance of excitation and inhibition and reveal insights into brain disorders (as mentioned above). Understanding how Exc and Inh neurons contribute to the overall network activity will enable researchers to create detailed models of how neural circuits function. This can lead to a deeper comprehension of complex behaviors and cognitive processes.

Materials and Methods

Visual Stimuli

All stimuli used in this study were previously used in the Kar et al. 2019[16] study and can be accessed at https://github.com/kohitij-kar/image_metrics. For a brief description of the stimuli, please refer below.

Generation of synthetic ("naturalistic") images:

High-quality images of single objects were generated using free ray-tracing software (http://www.povray.org), similar to Majaj et al. [2]. Each image consisted of a 2D projection of a 3D model (purchased from Dosch Design and TurboSquid) added to a random background. The eight objects chosen were **bear**, **elephant**, **face**, **apple**, **car**, **dog**, **chair**, and **plane**. By varying six viewing parameters, namely position (*x* and *y*), rotation (*x*, *y*, and *z*), and size, we explored three types of identity-preserving object variation. For each object, we generated 80 images. Thus, we generated a total of 640 images. All images were achromatic with a native resolution of 256 × 256 pixels.

Subjects

Four adult male rhesus macaques (Macaca mulatta) participated in our experiments. 3 of these macaques (monkeys N, B, and O, age ~ 10 years) contributed to the neural data analyzed here. A partially overlapping set of animals (monkeys M, N and B, age ~ 10 years) also provided behavioral data. All data were collected, and surgical and animal procedures were performed, in accordance with NIH guidelines, the Massachusetts Institute of Technology Committee on Animal Care, and the guidelines of Canadian Council on Animal Care on the use of laboratory animals and were also approved by the York University Animal Care Committee.

Primate behavioral testing

Active binary object discrimination task

We measured monkey behavior from 3 male rhesus macaques. Images were presented on a 24-inch LCD monitor (1920 × 1080 at 60 Hz) positioned 42.5 cm in front of the animal. Monkeys were head fixed. Monkeys fixated a white dot (0.2°) for 300 ms to initiate a trial. The trial started with the presentation of a sample image (from a set of 640 images) for 100 ms. This was followed by a blank gray screen for 100 ms, after which the choice screen was shown containing a standard image of the target object (the correct choice) and a standard image of the distractor object. The monkey was allowed to view freely the choice objects for up to 1500 ms and indicated its final choice by holding fixation over the selected object for 400 ms. Trials were aborted if gaze was not held within ±2° of the central fixation dot during any point until the choice screen was shown.

Prior to testing in the laboratory, monkeys were trained in their home-cages[11,58] to perform the delayed match to sample tasks on the same object categories (but with a different set of images).

Passive Fixation Task

During the passive viewing task, monkeys fixated a white dot (0.2°) for 300 ms to initiate a trial. We then presented a sequence of 5 to 10 images, each ON for 100 ms followed by a 100 ms gray (background) blank screen. This was followed by fluid (water) reward and an inter trial interval of 500 ms, followed by the next sequence. All neural data analyzed for this study was collected when the animals (n=3) were performing the passive fixation tasks.

Eye Tracking

We monitored eye movements using video eye tracking (SR Research EyeLink 1000). Using operant conditioning and water reward, our 4 subjects were trained to fixate a central white square (0.2°) within a square fixation window that ranged from ±2°. At the start of each behavioral session, monkeys performed an eye-tracking calibration task by making a saccade to a range of spatial targets and maintaining fixation for 500 ms. Calibration was repeated if drift was noticed over the course of the session.

Real-time eye-tracking was employed to ensure that eye jitter did not exceed ±2°, otherwise the trial was aborted, and data discarded. Stimulus display and reward control were managed using the MWorks Software (https://mworks.github.io).

Electrophysiological recording and data preprocessing

Prior to any neural recordings, we surgically implanted each monkey with a titanium headpost under aseptic conditions. We then chronically implanted 96-electrode Utah arrays (Blackrock Microsystems) in the IT cortex of the three monkeys: three each in the right cerebral hemisphere of monkey B, two in the right hemisphere of monkey N and three in the left hemisphere of monkey O . Each array consisted of a 4.4mm x 4.2mm square grid of electrodes spaced 400um apart. Each electrode (silicon shanks) in the array was 1.5mm long, with platinum tips. The arrays were pneumatically inserted and placed inferior to the superior temporal sulcus and anterior to the posterior middle temporal sulcus. We have previously analyzed multi-unit activity from these same monkeys with a significantly higher yield of reliable neural sites[16,17,49]. However, to obtain well-isolated single-units we made conservative choices (see below), for this study, that restricted our analyses to a total of 161 single neurons (N:64, B: 72, O:25) pooled across the three monkeys.

During each experimental session, multiunit neural activity was recorded continuously at a sampling rate of 20 kHz using an Intan RHD Recording Controller (Intan Technologies, LLC). The raw voltage signals were bandpass filtered offline using a second-order elliptical filter (300 Hz to 6 kHz, 0.1 dB passband ripple, 50 dB stopband attenuation), before being thresholded to obtain the multiunit spike events. A multiunit spike event was defined as the threshold crossing when voltage (falling edge) deviated by more than three times the standard deviation of the raw voltage values. For each multiunit spike event, we saved its corresponding waveform of length

1.6ms, with a pre-trough period of 0.4ms. These waveforms were then used to sort the multiunit spikes into individual neurons and for classification by putative cell-type.

Sorting multiunit neurons to single neurons

We normalized the 1.6ms long snippets of the multiunit spike waveforms to be between -1 and 1. Then, for each multiunit site, we projected the normalized waveforms into a lower dimensional space using principal components analysis (PCA) and used the top 2 principal components for k-means clustering. Each resulting waveform cluster was taken to belong to an individual neuron. To maximize the likelihood of selecting a neuron (and not a noisy signal), we computed the image rank order reliability (across splits of trial repetitions) for each neuron. For each neuron, we split the responses into half across repeated presentations of the same image, and then computed the Spearman-Brown corrected Pearson correlation between the two splits across images. We only retained neurons with internal reliability greater than 0.4.

Neuron cell-type classification based on spike waveform

For each neuron, we computed two features of its average action potential (AP), namely trough-to-peak time (or spike width) and repolarization time. Trough-to-peak is the time interval between the global minimum of AP to the subsequent positive peak. Repolarization time is the time interval between the global minimum of AP to baseline. Several past studies have observed a bimodal distribution of trough-to-peak time for excitatory and inhibitory neurons [8,25,28,34,59], and many have used it, or other correlated features of the AP, for classification of neurons into putative types [8,25,34,35,37–39,59–61] We discarded cells with trough-to-peak time greater than 1 ms or repolarization time greater than 0.6 s as these represent the tail ends of their respective distributions. We then used the k-means clustering algorithm on these two features to classify the cells into putative excitatory and putative inhibitory neurons.

Data Analyses

Behavioral Metrics (B. I_1)

We have used a one-vs-all image level behavioral performance metric ($B.I_1$ similar to the one used in [11,16]) to quantify the behavioral performance of the monkeys (described below). This metric estimates the overall discriminability of each image containing a specific target object from all other objects (pooling across all 7 possible distractor choices).

Given an image of object 'i', and all nine distractor objects ($j \neq i$) we computed the average performance per image as,

$$Performance^{i}_{image} = \frac{\sum_{j=1}^{8} Pc^{i,j\neq i}_{image}}{7}$$, where Pc refers to the fraction of correct responses for the binary task between objects 'i' and 'j'.

To compute the reliability of this vector, we split the trials per image into two equal halves by resampling without substitution. The median of the Spearman-Brown corrected correlation of the two corresponding vectors (one from each split half), across 1000 repetitions of the resampling was then used as the reliability score (i.e., internal consistency).

Neural recording metrics per IT neuron

Baseline Firing Rates

Baseline firing rates were estimated by counting the number of spikes evoked by gray (blank screen) images in the 70-170 ms time period after Test image onset and dividing by 100 ms.

Maximum Evoked Firing Rates

Maximum evoked firing rates per neuron were estimated by counting the number of spikes evoked by an image stimuli that produced the maximum response in the 70-170 ms time period after Test image onset and dividing by 100 ms.

Image rank-order response reliability per neuron

To estimate the reliability of the responses per neuron, we computed a Spearman-Brown corrected, split half (trial-based) correlation between the rank order of the image responses (across all 640 images).

Response Variance across images

Response variance was estimated per neuron as the variance in the spike count across 640 images in the 70-170 ms time period after Test image onset.

Category Selectivity

Category selectivity[44] measured the difference in average firing rate for each neuron between pairs of stimuli in different object categories [a between-category difference (BCD)] and the difference in activity between pairs of stimuli in the same category [a within-category difference (WCD)]. It was defined as the difference between BCD and WCD divided by their sum.

Estimation of IT population decode accuracies (N. I_1)

To estimate what information downstream neurons could easily "read" from a given IT neural population, we used a simple, biologically plausible linear decoder (i.e., linear classifiers), that has been previously shown to link IT population activity and primate behavior[2]. Such decoders are simple in that they can perform binary classifications by computing weighted sums (each weight is analogous to the strength of synapse) of input features and separate the outputs based on a decision boundary (analogous to a neuron's spiking threshold). Here we have used a logistic regression-based classifier. The learning model generates a decoder with a decision boundary that is optimized to best separate images of the target object from images of the distractor objects. The optimization is done under a regularization constraint that limits the complexity of the boundary. We used L2 (ridge) regularization, where the objective function for the minimization comprises of an additional term (to reduce model complexity),

$$\text{L2 (penalty)} = \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2$$

where β and p are the classifier weights associated with 'p' predictors (neurons). A stochastic gradient descent solver was used to estimate 8 (one for each object) one-vs-all classifiers. After training each of these classifiers with a 10-fold cross validation, we generated a class probability score (sc) per classifier for all held out test images given by,

$$sc = R\beta + bias,$$

where R is the population response vector, and the bias is estimated by the solver. The train and test sets were pseudo-randomly chosen multiple times until we every image of our image set was part of the held-out test set. We then computed the binary task performances, by calculating the percent correct score for each pair of possible binary task given an image. For instance, if an image was from object *i* (Target), then the percent correct score for the binary task between object i and object j (Distractor), $Pr^{i,j}$ was computed as,

$$Pr_{image}^{i,j} = \frac{sc_{image}^{i}}{sc_{image}^{i} + sc_{image}^{j}}$$

From each percent correct score, we then estimated a neural $N.I_1$ score (per image), following the same procedures as the behavioral metric.

Prediction of neural response from ANN features

We modeled each IT neural site as a linear combination of the ANN model features. We first extracted the features activations per image, from the ANN's layers. Using a 50%/50% train/test split of the images, we then estimated the regression weights (i.e., how we can linearly combine the model features to predict the neural site's responses) using a partial least squares regression procedure, using 25 retained components. The neural responses used for training (R^{TRAIN}) and testing (R^{TEST}) the encoding models were averaged firing rates (measured at the specific sites) within the time window 70 to 170 ms post stimulus onset. The training images used for regressing

the model features onto a neuron, at each time point, were sampled randomly (repeats included random subsampling) from the entire image set. For each set of regression weights (w) estimated on the training image responses (R^{TRAIN}), we generated the output of that 'synthetic neuron' for the held-out test set (M^{PRED}) as,

$$M^{PRED} = (w * F^{TEST}) + \beta,$$

where w and β are estimated via the PLS regression and F^{TEST} are the model activation features for the test image-set.

The percentage of explained variance, *IT predictivity* for that neural site, was then computed by normalizing the r^2 prediction value for that site by the self-consistency of the test image responses ($\rho^{R^{TEST}}$) for that site and the self-consistency of the regression model predictions ($\rho^{M^{PRED}}$) for that site (estimated by a Spearman Brown corrected trial-split correlation score).

$$\text{IT predictivity} = \left(\frac{corr(R^{TEST}, M^{PRED})}{\sqrt{\rho^{R^{TEST}} * \rho^{M^{PRED}}}}\right)^2$$

To achieve accurate cross-validation results, we had to test the prediction of the model on held out image responses.

Statistical Tests to determine significance of effects reported

We have used a combination of parametric and non-parametric tests in this study. Where parametric tests have been used, normal distributions of variables were ensured by the Lilliefors test. We have provided the test details and statistics within the manuscript text. For non-parametric tests, we employed permutation tests, where a null distribution was first obtained by randomly shuffling the labels of the two variables under comparison.

References

1. Tanaka, K. (1996). Inferotemporal Cortex and Object Vision. Annu. Rev. Neurosci. *19*, 109–139. 10.1146/annurev.ne.19.030196.000545.

2. Majaj, N.J., Hong, H., Solomon, E.A., and DiCarlo, J.J. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. J. Neurosci. *35*, 13402–13418. 10.1523/JNEUROSCI.5181-14.2015.

3. Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. Science *310*, 863–866. 10.1126/science.1117593.

4. Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., and Wu, C. (2004). Interneurons of the neocortical inhibitory system. Nat Rev Neurosci *5*, 793–807. 10.1038/nrn1519.

5. Hendry, S., Schwark, H., Jones, E., and Yan, J. (1987). Numbers and proportions of GABA-immunoreactive neurons in different areas of monkey cerebral cortex. J. Neurosci. *7*, 1503–1519. 10.1523/JNEUROSCI.07-05-01503.1987.

6. Tanigawa, H., Fujita, I., Kato, M., and Ojima, H. (1998). Distribution, morphology, and γ-aminobutyric acid immunoreactivity of horizontally projecting neurons in the macaque inferior temporal cortex. J. Comp. Neurol. *401*, 129–143. 10.1002/(SICI)1096-9861(19981109)401:1<129::AID-CNE8>3.0.CO;2-D.

7. Tamura, H., Kaneko, H., Kawasaki, K., and Fujita, I. (2004). Presumed Inhibitory Neurons in the Macaque Inferior Temporal Cortex: Visual Response Properties and Functional Interactions With Adjacent Neurons. Journal of Neurophysiology *91*, 2782–2796. 10.1152/jn.01267.2003.

8. Mruczek, R.E.B., and Sheinberg, D.L. (2012). Stimulus selectivity and response latency in putative inhibitory and excitatory neurons of the primate inferior temporal cortex. Journal of Neurophysiology *108*, 2725–2736. 10.1152/jn.00618.2012.

9. Wang, Y., Fujita, I., and Murayama, Y. (2000). Neuronal mechanisms of selectivity for object features revealed by blocking inhibition in inferotemporal cortex. Nat Neurosci *3*, 807–813. 10.1038/77712.

10. Yamins, D.L.K., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci *19*, 356–365. 10.1038/nn.4244.

11. Rajalingham, R., Issa, E.B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J.J. (2018). Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. J Neurosci *38*, 7255–7269. 10.1523/JNEUROSCI.0388-18.2018.

12. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., and Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv:1811.12231 [cs, q-bio, stat].

13. Goodfellow, I.J., Shlens, J., and Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. arXiv:1412.6572 [cs, stat].

14. Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? (Neuroscience) 10.1101/407007.

15. Kubilius, J., Schrimpf, M., Kar, K., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. arXiv:1909.06161 [cs, eess, q-bio].

16. Kar, K., Kubilius, J., Schmidt, K., Issa, E.B., and DiCarlo, J.J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat Neurosci *22*, 974–983. 10.1038/s41593-019-0392-5.

17. Kar, K., and DiCarlo, J.J. (2021). Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition. Neuron *109*, 164-176.e5. 10.1016/j.neuron.2020.09.035.

18. Bao, P., She, L., McGill, M., and Tsao, D.Y. (2020). A map of object space in primate inferotemporal cortex. Nature *583*, 103–108. 10.1038/s41586-020-2350-5.

19. 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 (Neuroscience) 10.1101/2020.07.09.185116.

20. Margalit, E., Lee, H., Finzi, D., DiCarlo, J.J., Grill-Spector, K., and Yamins, D.L.K. (2023). A Unifying Principle for the Functional Organization of Visual Cortex (Neuroscience) 10.1101/2023.05.18.541361.

21. Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M.C., DiCarlo, J.J., and Yamins, D.L.K. (2021). Unsupervised neural network models of the ventral visual stream. Proc. Natl. Acad. Sci. U.S.A. *118*, e2014196118. 10.1073/pnas.2014196118.

22. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A *111*, 8619–8624. 10.1073/pnas.1403112111.

23. Mehrer, J., Spoerer, C.J., Jones, E.C., Kriegeskorte, N., and Kietzmann, T.C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. Proc. Natl. Acad. Sci. U.S.A. *118*, e2011417118. 10.1073/pnas.2011417118.

24. Dobs, K., Martinez, J., Kell, A.J.E., and Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. Sci. Adv. *8*, eabl8913. 10.1126/sciadv.abl8913.

25. McCormick, D.A., Connors, B.W., Lighthall, J.W., and Prince, D.A. (1985). Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. Journal of Neurophysiology *54*, 782–806. 10.1152/jn.1985.54.4.782.

26. Connors, B.W., and Gutnick, M.J. (1990). Intrinsic firing patterns of diverse neocortical neurons. Trends in Neurosciences *13*, 99–104. 10.1016/0166-2236(90)90185-D.

27. Kawaguchi, Y., and Kubota, Y. (1993). Correlation of physiological subgroupings of nonpyramidal cells with parvalbumin- and calbindinD28k-immunoreactive neurons in layer V of rat frontal cortex. Journal of Neurophysiology *70*, 387–396. 10.1152/jn.1993.70.1.387.

28. Barthó, P., Hirase, H., Monconduit, L., Zugaro, M., Harris, K.D., and Buzsáki, G. (2004). Characterization of Neocortical Principal Cells and Interneurons by Network Interactions and Extracellular Features. Journal of Neurophysiology *92*, 600–608. 10.1152/jn.01170.2003.

29. Simons, D.J. (1978). Response properties of vibrissa units in rat SI somatosensory neocortex. Journal of Neurophysiology *41*, 798–820. 10.1152/jn.1978.41.3.798.

30. Mountcastle, V.B., Talbot, W.H., Sakata, H., and Hyvärinen, J. (1969). Cortical neuronal mechanisms in flutter-vibration studied in unanesthetized monkeys. Neuronal periodicity and frequency discrimination. Journal of Neurophysiology *32*, 452–484. 10.1152/jn.1969.32.3.452.

31. Gibson, J.R., Beierlein, M., and Connors, B.W. (1999). Two networks of electrically coupled inhibitory neurons in neocortex. Nature *402*, 75–79. 10.1038/47035.

32. Dykes, R.W., Lamour, Y., Diadori, P., Landry, P., and Dutar, P. (1988). Somatosensory cortical neurons with an identifiable electrophysiological signature. Brain Research *441*, 48–58. 10.1016/0006-8993(88)91382-0.

33. Gray, C.M., and McCormick, D.A. (1996). Chattering Cells: Superficial Pyramidal Neurons Contributing to the Generation of Synchronous Oscillations in the Visual Cortex. Science *274*, 109–113. 10.1126/science.274.5284.109.

34. Mitchell, J.F., Sundberg, K.A., and Reynolds, J.H. (2007). Differential Attention-Dependent Response Modulation across Cell Classes in Macaque Visual Area V4. Neuron *55*, 131–141. 10.1016/j.neuron.2007.06.018.

35. Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J.J. (2007). Trade-Off between Object Selectivity and Tolerance in Monkey Inferotemporal Cortex. J. Neurosci. *27*, 12292–12307. 10.1523/JNEUROSCI.1897-07.2007.

36. Wilson, F.A., O'Scalaidhe, S.P., and Goldman-Rakic, P.S. (1994). Functional synergism between putative gamma-aminobutyrate-containing neurons and pyramidal neurons in prefrontal cortex. Proc. Natl. Acad. Sci. U.S.A. *91*, 4009–4013. 10.1073/pnas.91.9.4009.

37. Rao, S.G., Williams, G.V., and Goldman-Rakic, P.S. (1999). Isodirectional Tuning of Adjacent Interneurons and Pyramidal Cells During Working Memory: Evidence for Microcolumnar Organization in PFC. Journal of Neurophysiology *81*, 1903–1916. 10.1152/jn.1999.81.4.1903.

38. Constantinidis, C., and Goldman-Rakic, P.S. (2002). Correlated Discharges Among Putative Pyramidal Neurons and Interneurons in the Primate Prefrontal Cortex. Journal of Neurophysiology *88*, 3487–3497. 10.1152/jn.00188.2002.

39. González-Burgos, G., Krimer, L.S., Povysheva, N.V., Barrionuevo, G., and Lewis, D.A. (2005). Functional Properties of Fast Spiking Interneurons and Their Synaptic Connections With Pyramidal Cells in Primate Dorsolateral Prefrontal Cortex. Journal of Neurophysiology *93*, 942–953. 10.1152/jn.00787.2004.

40. Kerlin, A.M., Andermann, M.L., Berezovskii, V.K., and Reid, R.C. (2010). Broadly Tuned Response Properties of Diverse Inhibitory Neuron Subtypes in Mouse Visual Cortex. Neuron *67*, 858–871. 10.1016/j.neuron.2010.08.002.

41. Cardin, J.A., Palmer, L.A., and Contreras, D. (2007). Stimulus Feature Selectivity in Excitatory and Inhibitory Neurons in Primary Visual Cortex. J. Neurosci. *27*, 10333–10344. 10.1523/JNEUROSCI.1692-07.2007.

42. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds. (Curran Associates, Inc.).

43. He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition.

44. Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. (2001). Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. Science *291*, 312–316. 10.1126/science.291.5502.312.

45. Divac, I., Rosvold, H.E., and Szwarcbart, M.K. (1967). Behavioral effects of selective ablation of the caudate nucleus. Journal of Comparative and Physiological Psychology *63*, 184–190. 10.1037/h0024348.

46. Webster, M., Ungerleider, L., and Bachevalier, J. (1991). Connections of inferior temporal areas TE and TEO with medial temporal- lobe structures in infant and adult monkeys. J. Neurosci. *11*, 1095–1116. 10.1523/JNEUROSCI.11-04-01095.1991.

47. Wang, Y. (2002). Contribution of GABAergic Inhibition to Receptive Field Structures of Monkey Inferior Temporal Neurons. Cerebral Cortex *12*, 62–74. 10.1093/cercor/12.1.62.

48. Li, P., Cornford, J., Ghosh, A., and Richards, B. (2023). Learning better with Dale's Law: A Spectral Perspective (Neuroscience) 10.1101/2023.06.28.546924.

49. Dapello, J., Kar, K., Schrimpf, M., Geary, R., Ferguson, M., Cox, D.D., and DiCarlo, J.J. (2022). Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness (Neuroscience) 10.1101/2022.07.01.498495.

50. Rubenstein, J.L.R., and Merzenich, M.M. (2003). Model of autism: increased ratio of excitation/inhibition in key neural systems: **Model of autism**. Genes, Brain and Behavior *2*, 255–267. 10.1034/j.1601-183X.2003.00037.x.

51. Maestú, F., De Haan, W., Busche, M.A., and DeFelipe, J. (2021). Neuronal excitation/inhibition imbalance: core element of a translational perspective on Alzheimer pathophysiology. Ageing Research Reviews *69*, 101372. 10.1016/j.arr.2021.101372.

52. Thye, M.D., Bednarz, H.M., Herringshaw, A.J., Sartin, E.B., and Kana, R.K. (2018). The impact of atypical sensory processing on social impairments in autism spectrum disorder. Developmental Cognitive Neuroscience *29*, 151–167. 10.1016/j.dcn.2017.04.010.

53. Bijlenga, D., Tjon-Ka-Jie, J.Y.M., Schuijers, F., and Kooij, J.J.S. (2017). Atypical sensory profiles as core features of adult ADHD, irrespective of autistic symptoms. Eur. psychiatr. *43*, 51–57. 10.1016/j.eurpsy.2017.02.481.

54. Kar, K. (2022). A Computational Probe into the Behavioral and Neural Markers of Atypical Facial Emotion Processing in Autism. J. Neurosci. *42*, 5115–5126. 10.1523/JNEUROSCI.2229-21.2022.

55. Trainito, C., Von Nicolai, C., Miller, E.K., and Siegel, M. (2019). Extracellular Spike Waveform Dissociates Four Functionally Distinct Cell Classes in Primate Cortex. Current Biology *29*, 2973-2982.e5. 10.1016/j.cub.2019.07.051.

56. Rajalingham, R., Sorenson, M., Azadi, R., Bohn, S., DiCarlo, J.J., and Afraz, A. (2021). Chronically implantable LED arrays for behavioral optogenetics in primates. Nat Methods *18*, 1112–1116. 10.1038/s41592-021-01238-9.

57. Roth, B.L. (2016). DREADDs for Neuroscientists. Neuron *89*, 683–694. 10.1016/j.neuron.2016.01.040.

58. Giverin, C., Ramezanpour, H., and Kar, K. (2023). Low-cost, portable, easy-to-use kiosks to facilitate home-cage testing of non-human primates during vision-based behavioral tasks (Open Science Framework) 10.31219/osf.io/csdzv.

59. Woloszyn, L., and Sheinberg, D.L. (2012). Effects of Long-Term Visual Experience on Responses of Distinct Classes of Single Units in Inferior Temporal Cortex. Neuron *74*, 193–205. 10.1016/j.neuron.2012.01.032.

60. Bruno, R.M., and Simons, D.J. (2002). Feedforward Mechanisms of Excitatory and Inhibitory Cortical Receptive Fields. J. Neurosci. *22*, 10966–10975. 10.1523/JNEUROSCI.22-24-10966.2002.

61. Trainito, C., Von Nicolai, C., Miller, E.K., and Siegel, M. (2019). Extracellular Spike Waveform Dissociates Four Functionally Distinct Cell Classes in Primate Cortex. Current Biology *29*, 2973-2982.e5. 10.1016/j.cub.2019.07.051.

Supplementary Material

Model name	Layer name
densenet-201	pool4_pool
densenet-169	pool4_pool
resnet-152_v1	resnet_v1_152/block3/unit_36/bottleneck_v1
resnet-101_v1	resnet_v1_101/block3/unit_23/bottleneck_v1
mobilenet_v1_0.75_224	Conv2d_12_depthwise
densenet-121	pool4_pool
resnet50-SIN_IN_IN	layer4.0.relu
CORnet-S	IT.output-t0
nasnet_large	Cell_12
inception_v3	Mixed_7a
resnet-50_v2	resnet_v2_50/block4/unit_1/bottleneck_v2
inception_v2	Mixed_5a
resnet-152_v2	resnet_v2_152/block4/unit_1/bottleneck_v2
pnasnet_large	Cell_8
alexnet	feature.12
squeezenet1_1	features.9.expand3x3_activation
resnext101_32x8d_wsl	layer3.3.relu
squeezenet1_0	features.12.expand3x3_activation
mobilenet_v1_0.25_128	Conv2d_11_depthwise

Figure S1. **List of all ANNs.** We picked a representative sample of 19 ANNs from Brain-Score. All models on Brain-Score have their layers 'committed' to ventral visual stream areas V1, V2, V4 and IT, that is each model has its own V1, V2, V4 and IT layers defined, even though these might not have been explicitly labeled as so by the model architects. Layer selection procedure is run using separate public datasets (see https://github.com/brain-score/brain-score/issues/303 for additional details). The layer names listed here are the IT layers for the corresponding models. All names here match the identifiers used by Brain-Score model commitments.