Performance-optimized hierarchical models predict neural responses in higher visual cortex

Daniel L. K. Yamins[a,1], Ha Hong[a,b,1], Charles F. Cadieu[a], Ethan A. Solomon[a], Darren Seibert[a], and James J. DiCarlo[a,2]

[a]Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Harvard-MIT Division of Health Sciences and Technology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139

The ventral visual stream underlies key human visual object recognition abilities. However, neural encoding in the higher areas of the ventral stream remains poorly understood. Here, we describe a modeling approach that yields a quantitatively accurate model of inferior temporal (IT) cortex, the highest ventral cortical area. Using high-throughput computational techniques, we discovered that, within a class of biologically plausible hierarchical neural network models, there is a strong correlation between a model's categorization performance and its ability to predict individual IT neural unit response data. To pursue this idea, we then identified a high-performing neural network that matches human performance on a range of recognition tasks. Critically, even though we did not constrain this model to match neural data, its top output layer turns out to be highly predictive of IT spiking responses to complex naturalistic images at both the single site and population levels. Moreover, the model's intermediate layers are highly predictive of neural responses in the V4 cortex, a midlevel visual area that provides the dominant cortical input to IT. These results show that performance optimization—applied in a biologically appropriate model class—can be used to build quantitative predictive models of neural processing.

computational neuroscience | computer vision | array electrophysiology

Retinal images of real-world objects vary drastically due to changes in object pose, size, position, lighting, nonrigid deformation, occlusion, and many other sources of noise and variation. Humans effortlessly recognize objects rapidly and accurately despite this enormous variation, an impressive computational feat (1). This ability is supported by a set of interconnected brain areas collectively called the ventral visual stream (2, 3), with homologous areas in nonhuman primates (4). The ventral stream is thought to function as a series of hierarchical processing stages (5–7) that encode image content (e.g., object identity and category) increasingly explicitly in successive cortical areas (1, 8, 9). For example, neurons in the lowest area, V1, are well described by Gabor-like edge detectors that extract rough object outlines (10), although the V1 population does not show robust tolerance to complex image transformations (9). Conversely, rapidly evoked population activity in top-level inferior temporal (IT) cortex can directly support real-time, invariant object categorization over a wide range of tasks (11, 12). Midlevel ventral areas—such as V4, the dominant cortical input to IT—exhibit intermediate levels of object selectivity and variation tolerance (12–14).

Significant progress has been made in understanding lower ventral areas such as V1, where conceptually compelling models have been discovered (10). These models are also quantitatively accurate and can predict response magnitudes of individual neuronal units to novel image stimuli. Higher ventral cortical areas, especially V4 and IT, have been much more difficult to understand. Although first principles-based models of higher ventral cortex have been proposed (15–20), these models fail to match important features of the higher ventral visual neural representation in both humans and macaques (4, 21). Moreover, attempts to fit V4 and IT neural tuning curves on general image stimuli have shown only limited predictive success (22, 23).

Explaining the neural encoding in these higher ventral areas thus remains a fundamental open question in systems neuroscience.

As with V1, models of higher ventral areas should be neurally predictive. However, because the higher ventral stream is also believed to underlie sophisticated behavioral object recognition capacities, models must also match IT on performance metrics, equalling (or exceeding) the decoding capacity of IT neurons on object recognition tasks. A model with perfect neural predictivity in IT will necessarily exhibit high performance, because IT itself does. Here we demonstrate that the converse is also true, within a biologically appropriate model class. Combining high-throughput computational and electrophysiology techniques, we explore a wide range of biologically plausible hierarchical neural network models and then assess them against measured IT and V4 neural response data. We show that there is a strong correlation between a model's performance on a challenging high-variation object recognition task and its ability to predict individual IT neural unit responses.

Extending this idea, we used optimization methods to identify a neural network model that matches human performance on a range of recognition tasks. We then show that even though this model was never explicitly constrained to match neural data, its output layer is highly predictive of neural responses in IT cortex—providing a first quantitatively accurate model of this highest ventral cortex area. Moreover, the middle layers of the model are highly predictive of V4 neural responses, suggesting top-down performance constraints directly shape intermediate visual representations.

Significance

Humans and monkeys easily recognize objects in scenes. This ability is known to be supported by a network of hierarchically interconnected brain areas. However, understanding neurons in higher levels of this hierarchy has long remained a major challenge in visual systems neuroscience. We use computational techniques to identify a neural network model that matches human performance on challenging object categorization tasks. Although not explicitly constrained to match neural data, this model turns out to be highly predictive of neural responses in both the V4 and inferior temporal cortex, the top two layers of the ventral visual hierarchy. In addition to yielding greatly improved models of visual cortex, these results suggest that a process of biological performance optimization directly shaped neural mechanisms.

Results

Invariant Object Recognition Performance Strongly Correlates with IT Neural Predictivity. We first measured IT neural responses on a benchmark testing image set that exposes key performance characteristics of visual representations (24). This image set consists of 5,760 images of photorealistic 3D objects drawn from eight natural categories (animals, boats, cars, chairs, faces, fruits, planes, and tables) and contains high levels of the object position, scale, and pose variation that make recognition difficult for artificial vision systems, but to which humans are robustly tolerant (1, 25). The objects are placed on cluttered natural scenes that are randomly selected to ensure background content is uncorrelated with object identity (Fig. S1A).

Using multiple electrode arrays, we collected responses from 168 IT neurons to each image. We then used high-throughput computational methods to evaluate thousands of candidate neural network models on these same images, measuring object categorization performance as well as IT neural predictivity for each model (Fig. 1A; each point represents a distinct model). To measure categorization performance, we trained support vector machine (SVM) linear classifiers on model output layer units (11) and computed cross-validated testing accuracy for these trained classifiers. To assess models' neural predictivity, we used a standard linear regression methodology (10, 26, 27): for each target IT neural site, we identified a synthetic neuron composed of a linear weighting of model outputs that would best match that site on fixed sample images and then tested response predictions against actual neural site's output on novel images (*Materials and Methods* and *SI Text*).

Models were drawn from a large parameter space of convolutional neural networks (CNNs) expressing an inclusive version of the hierarchical processing concept (17, 18, 20, 28). CNNs approximate the general retinotopic organization of the ventral stream via spatial convolution, with computations in any one region of the visual field identical to those elsewhere. Each convolutional layer is composed of simple and neuronally plausible basic operations, including linear filtering, thresholding, pooling, and normalization (Fig. S2A). These layers are stacked hierarchically to construct deep neural networks.

Each model is specified by a set of 57 parameters controlling the number of layers and parameters at each layer, fan-in and fan-out, activation thresholds, pooling exponents, and local receptive field sizes at each layer. Network depth ranged from one to three layers, and filter weights for each layer were chosen randomly from bounded uniform distributions whose bounds were model parameters (*SI Text*). These models are consistent with the Hierarchical Linear-Nonlinear (HLN) hypothesis that higher level neurons (e.g., IT) output a linear weighting of inputs from intermediate-level (e.g., V4) neurons followed by simple additional nonlinearities (14, 16, 29).

Models were selected for evaluation by one of three procedures: (*i*) random sampling of the uniform distribution over parameter space (Fig. 1A; n = 2,016, green dots); (*ii*) optimization for performance on the high-variation eight-way categorization task (n = 2,043, blue dots); and (*iii*) optimization directly for IT neural predictivity (n = 1,876, orange dots; also see *SI Text* and Fig. S3). In each case, we observed significant variation in both performance and IT predictivity across the parameter range. Thus, although the HLN hypothesis is consistent with a broad spectrum of particular neural network architectures, specific parameter choices have a large effect on a given model's recognition performance and neural predictivity.

Performance was significantly correlated with neural predictivity in all three selection regimes. Models that performed better on the categorization task were also more likely to produce outputs more closely aligned to IT neural responses. Although the class of HLN-consistent architectures contains many neurally inconsistent architectures with low IT predictivity, performance provides a meaningful way to a priori rule out many of those inconsistent models. No individual model parameters correlated nearly as strongly with IT predictivity as performance (Fig. S4), indicating that the performance/IT predictivity correlation cannot be explained by simpler mechanistic considerations (e.g., receptive field size of the top layer).

Critically, directed optimization for performance significantly increased the correlation with IT predictivity compared with the random selection regime (r = 0.78 vs. r = 0.55), even though neural data were not used in the optimization. Moreover, when optimizing for performance, the best-performing models predicted neural output as well as those models directly selected for neural predictivity, although the reverse is not true. Together, these results imply that, although the IT predictivity metric is a complex function of the model parameter landscape, performance optimization is an efficient means to identify regions in parameter space containing IT-like models.

IT Cortex as a Neural Performance Target. Fig. 1A suggests a next step toward improved encoding models of higher ventral cortex: drive models further to the right along the x axis—if the correlation holds, the models will also climb on the y axis. Ideally, this would involve identifying hierarchical neural networks that perform at or near human object recognition performance levels and validating them using rigorous tests against neural data (Fig. 2A). However, the difficulty of meeting the performance challenge itself can be seen in Fig. 2B. To obtain neural reference points on categorization performance, we trained linear

Fig. 1. Performance/IT-predictivity correlation. (*A*) Object categorization performance vs. IT neural explained variance percentage (IT-predictivity) for CNN models in three independent high-throughput computational experiments (each point is a distinct neural network architecture). The x axis shows performance (balanced accuracy, chance is 0.5) of the model output features on a high-variation categorization task; the y axis shows the median single site IT explained variance percentage (n = 168 sites) of that model. Each dot corresponds to a distinct model selected from a large family of convolutional neural network architectures. Models were selected by random draws from parameter space (green dots), object categorization performance-optimization (blue dots), or explicit IT predictivity optimization (orange dots). (*B*) Pursuing the correlation identified in *A*, a high-performing neural network was identified that matches human performance on a range of recognition tasks, the HMO model. The object categorization performance vs. IT neural predictivity correlation extends across a variety of models exhibiting a wide range of performance levels. Black circles include controls and published models; red squares are models produced during the HMO optimization procedure. The category ideal observer (purple square) lies significantly off the main trend, but is not an actual image-computable model. The r value is computed over red and black points. For reference, light blue circles indicate performance optimized models (blue dots) from *A*.
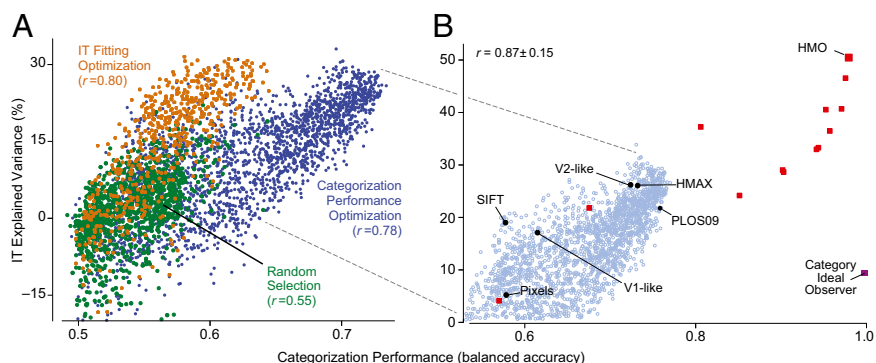
classifiers on the IT neural population (Fig. 2B, green bars) and the V4 neural population ($n = 128$, hatched green bars). To expose a key axis of recognition difficulty, we computed performance results at three levels of object view variation, from low (fixed orientation, size, and position) to high (180° rotations on all axes, 2.5× dilation, and full-frame translations; Fig. S1A). As a behavioral reference point, we also measured human performance on these tasks using web-based crowdsourcing methods (black bars). A crucial observation is that at all levels of variation, the IT population tracks human performance levels, consistent with known results about IT's high category decoding abilities (11, 12). The V4 population matches IT and human performance at low levels of variation, but performance drops quickly at higher variation levels. (This V4-to-IT performance gap remains nearly as large even for images with no object translation variation, showing that the performance gap is not due just to IT's larger receptive fields.)

As a computational reference, we used the same procedure to evaluate a variety of published ventral stream models targeting several levels of the ventral hierarchy. To control for low-level confounds, we tested the (trivial) pixel model, as well as SIFT, a simple baseline computer vision model (30). We also evaluated a V1-like Gabor-based model (25), a V2-like conjunction-of-Gabors model (31), and HMAX (17, 28), a model targeted at explaining higher ventral cortex and that has receptive field sizes
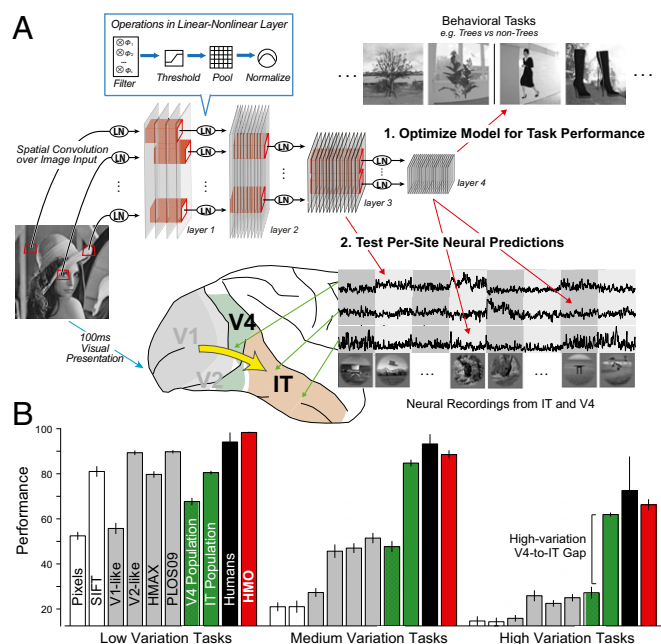

Fig. 2. Neural-like models via performance optimization. (A) We (1) used high-throughput computational methods to optimize the parameters of a hierarchical CNN with linear-nonlinear (LN) layers for performance on a challenging invariant object recognition task. Using new test images distinct from those used to optimize the model, we then (2) compared output of each of the model's layers to IT neural responses and the output of intermediate layers to V4 neural responses. To obtain neural data for comparison, we used chronically implanted multielectrode arrays to record the responses of multiunit sites in IT and V4, obtaining the mean visually evoked response of each of 296 neural sites to ~6,000 complex images. (B) Object categorization performance results on the test images for eight-way object categorization at three increasing levels of object view variation (y axis units are 8-way categorization percent-correct, chance is 12.5%). IT (green bars) and V4 (hatched green bars) neural responses, and computational models (gray and red bars) were collected on the same image set and used to train support vector machine (SVM) linear classifiers from which population performance accuracy was evaluated. Error bars are computed over train/test image splits. Human subject responses on the same tasks were collected via psychophysics experiments (black bars); error bars are due to intersubject variation.

similar to those observed in IT. The HMAX model can be trained in a domain-specific fashion, and to give it the best chance of success, we performed this training using the benchmark images themselves (see *SI Text* for more information on the comparison models). Like V4, the control models that we tested approach IT and human performance levels in the low-variation condition, but in the high-variation condition, all of them fail to match the performance of IT units by a large margin. It is not surprising that V1 and V2 models are not nearly as effective as IT, but it is instructive to note that the task is sufficiently difficult that the HMAX model performs less well than the V4 population sample, even when pretrained directly on the test dataset.

Constructing a High-Performing Model. Although simple three-layer hierarchical CNNs can be effective at low-variation object recognition tasks, recent work has shown that they may be limited in their performance capacity for higher-variation tasks (9). For this reason, we extended our model class to contain combinations (e.g., mixtures) of deeper CNN networks (Fig. S2B), which correspond intuitively to architecturally specialized subregions like those observed in the ventral visual stream (13, 32). To address the significant computational challenge of finding especially high-performing architectures within this large space of possible networks, we used hierarchical modular optimization (HMO). The HMO procedure embodies a conceptually simple hypothesis for how high-performing combinations of functionally specialized hierarchical architectures can be efficiently discovered and hierarchically combined, without needing to prespecify the subtasks ahead of time. Algorithmically, HMO is analogous to an adaptive boosting procedure (33) interleaved with hyperparameter optimization (see *SI Text* and Fig. S2C).

As a pretraining step, we applied the HMO selection procedure on a screening task (Fig. S1B). Like the testing set, the screening set contained images of objects placed on randomly selected backgrounds, but used entirely different objects in totally nonoverlapping semantic categories, with none of the same backgrounds and widely divergent lighting conditions and noise levels. Like any two samples of naturalistic images, the screening and testing images have high-level commonalities but quite different semantic content. For this reason, performance increases that transfer between them are likely to also transfer to other naturalistic image sets. Via this pretraining, the HMO procedure identified a four-layer CNN with 1,250 top-level outputs (Figs. S2B and S5), which we will refer to as the HMO model.

Using the same classifier training protocol as with the neural data and control models, we then tested the HMO model to determine whether its performance transferred from the screening to the testing image set. In fact, the HMO model matched the object recognition performance of the IT neural sample (Fig. 2B, red bars), even when faced with large amounts of variation—a hallmark of human object recognition ability (1). These performance results are robust to the number of training examples and number of sampled model neurons, across a variety of distinct recognition tasks (Figs. S6 and S7).

Predicting Neural Responses in Individual IT Neural Sites. Given that the HMO model had plausible performance characteristics, we then measured its IT predictivity, both for the top layer and each of the three intermediate layers (Fig. 3, red lines/bars). We found that each successive layer predicted IT units increasingly well, demonstrating that the trend identified in Fig. 1A continues to hold in higher performance regimes and across a wide range of model complexities (Fig. 1B). Qualitatively examining the specific predictions for individual images, the model layers show that category selectivity and tolerance to more drastic image transformations emerges gradually along the hierarchy (Fig. 3A, top four rows). At lower layers, model units predict IT responses only at a limited range of object poses and positions. At higher layers, variation tolerance grows while category selectivity develops, suggesting that as more explicit "untangled" object recognition

features are generated at each stage, the representations become increasingly IT-like (9).

Critically, we found that the top layer of the high-performing HMO model achieves high predictivity for individual IT neural sites, predicting $48.5 \pm 1.3\%$ of the explainable IT neuronal variance (Fig. 3 *B* and *C*). This represents a nearly 100% improvement over the best comparison models and is comparable to the prediction accuracy of state-of-the-art models of lower-level ventral areas such as V1 on complex stimuli (10). In comparison, although the HMAX model was better at predicting IT responses than baseline V1 or SIFT, it was not significantly different from the V2-like model.

To control for how much neural predictivity should be expected from any algorithm with high categorization performance, we assessed semantic ideal observers (34), including a hypothetical model that has perfect access to all category labels. The ideal observers do predict IT units above chance level (Fig. 3*C*, left two bars), consistent with the observation that IT neurons are partially categorical. However, the ideal observers are significantly less predictive than the HMO model, showing that high IT predictivity does not automatically follow from category selectivity and that there is significant noncategorical structure in IT responses attributable to intrinsic aspects of hierarchical network structure (Fig. 3*A*, last row). These results suggest that high categorization performance and the hierarchical model architecture class work in concert to produce IT-like populations, and neither of these constraints is sufficient on its own to do so.

Population Representation Similarity. Characterizing the IT neural representation at the population level may be equally important for understanding object visual representation as individual IT neural sites. The representation dissimilarity matrix (RDM) is a

convenient tool comparing two representations on a common stimulus set in a task-independent manner (4, 35). Each entry in the RDM corresponds to one stimulus pair, with high/low values indicating that the population as a whole treats the pair stimuli as very different/similar. Taken over the whole stimulus set, the RDM characterizes the layout of the images in the high-dimensional neural population space. When images are ordered by category, the RDM for the measured IT neural population (Fig. 4*A*) exhibits clear block-diagonal structure—associated with IT's exceptionally high categorization performance—as well as off-diagonal structure that characterizes the IT neural representation more finely than any single performance metric (Fig. 4*A* and Fig. S8). We found that the neural population predicted by the output layer of the HMO model had very high similarity to the actual IT population structure, close to the split-half noise ceiling of the IT population (Fig. 4*B*). This implies that much of the residual variance unexplained at the single-site level may not be relevant for object recognition in the IT population level code.

We also performed two stronger tests of generalization: (*i*) object-level generalization, in which the regressor training set contained images of only 32 object exemplars (four in each of eight categories), with RDMs assessed only on the remaining 32 objects, and (*ii*) category-level generalization, in which the regressor sample set contained images of only half the categories but RDMs were assessed only on images of the other categories (see Figs. S8 and S9). We found that the prediction generalizes robustly, capturing the IT population's layout for images of completely novel objects and categories (Fig. 4 *B* and *C* and Fig. S8).

Predicting Responses in V4 from Intermediate Model Layers. Cortical area V4 is the dominant cortical input to IT, and the neural representation in V4 is known to be significantly less categorical
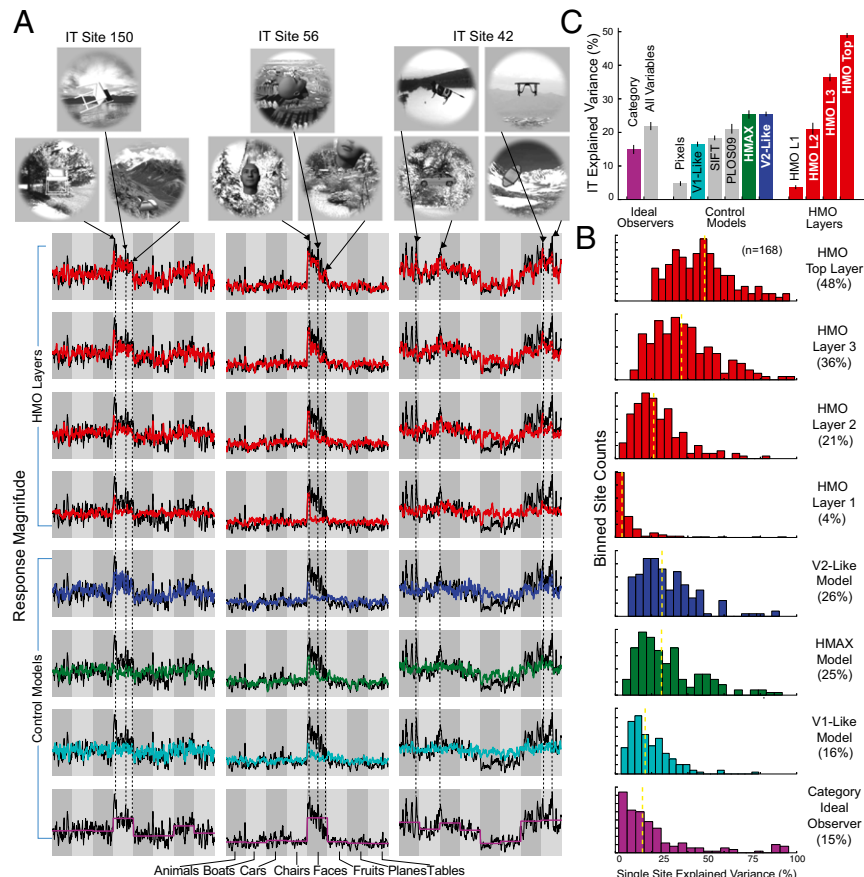

Fig. 3. IT neural predictions. (*A*) Actual neural response (black trace) vs. model predictions (colored trace) for three individual IT neural sites. The *x* axis in each plot shows 1,600 test images sorted first by category identity and then by variation amount, with more drastic image transformations toward the right within each category block. The *y* axis represents the prediction/response magnitude of the neural site for each test image (those not used to fit the model). Two of the units show selectivity for specific classes of objects, namely chairs (*Left*) and faces (*Center*), whereas the third (*Right*) exhibits a wider variety of image preferences. The four top rows show neural predictions using the visual feature set (i.e., units sampled) from each of the four layers of the HMO model, whereas the lower rows show the those of control models. (*B*) Distributions of model explained variance percentage, over the population of all measured IT sites (*n* = 168). Yellow dotted line indicates distribution median. (*C*) Comparison of IT neural explained variance percentage for various models. Bar height shows median explained variance, taken over all predicted IT units. Error bars are computed over image splits. Colored bars are those shown in *A* and *B*, whereas gray bars are additional comparisons.
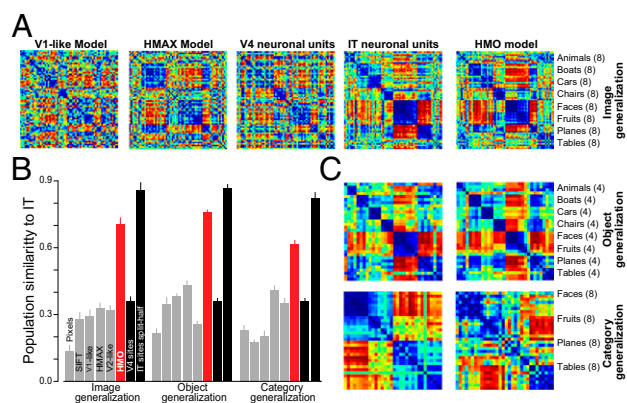
Fig. 4. Population-level similarity. (*A*) Object-level representation dissimilarity matrices (RDMs) visualized via rank-normalized color plots (blue = 0th distance percentile, red = 100th percentile). (*B*) IT population and the HMO-based IT model population, for image, object, and category generalizations (*SI Text*). (*C*) Quantification of model population representation similarity to IT. Bar height indicates the spearman correlation value of a given model's RDM to the RDM for the IT neural population. The IT bar represents the Spearman-Brown corrected consistency of the IT RDM for split-halves over the IT units, establishing a noise-limited upper bound. Error bars are taken over cross-validated regression splits in the case of models and over image and unit splits in the case of neural data.

than that of IT (12). Comparing a performance-optimized model to these data would provide a strong test both of its ability to predict the internal structure of the ventral stream, as well as to go beyond the direct consequences of category selectivity. We thus measured the HMO model's neural predictivity for the V4 neural population (Fig. 5). We found that the HMO model's penultimate layer is highly predictive of V4 neural responses (51.7 ± 2.3% explained V4 variance), providing a significantly better match to V4 than either the model's top or bottom layers. These results are strong evidence for the hypothesis that V4 corresponds to an intermediate layer in a hierarchical model whose top layer is an effective model of IT. Of the control models that we tested, the V2-like model predicts the most V4 variation (34.1 ± 2.4%). Unlike the case of IT, semantic models explain effectively no variance in V4, consistent with V4's lack of category selectivity. Together these results suggest that performance optimization not only drives top-level output model layers to resemble IT, but also imposes biologically consistent constraints on the intermediate feature representations that can support downstream performance.

Discussion

Here, we demonstrate a principled method for achieving greatly improved predictive models of neural responses in higher ventral cortex. Our approach operationalizes a hypothesis for how two biological constraints together shaped visual cortex: (*i*) the functional constraint of recognition performance and (*ii*) the structural constraint imposed by the hierarchical network architecture.

Generative Basis for Higher Visual Cortical Areas. Our modeling
approach has common ground with existing work on neural response prediction (27), e.g., the HLN hypothesis. However, in a departure from that line of work, we do not tune model parameters (the nonlinearities or the model filters) separately for each neural unit to be predicted. In fact, with the exception of the final linear weighting, we do not tune parameters using neural data at all. Instead, the parameters of our model were independently selected to optimize functional performance at the top level, and these choices create fixed bases from which any individual IT or V4 unit can be composed. This yields a generative model that allows the sampling of an arbitrary number of

neurally consistent units. As a result, the size of the model does not scale with the number of neural sites to be predicted—and because the prediction results were assessed for a random sample of IT and V4 units, they are likely to generalize with similar levels of predictivity to any new sites that are measured.

What Features Do Good Models Share? Although the highest-performing models had certain commonalities (e.g., more hierarchical layers), many poor models also exhibited these features, and no one architectural parameter dominated performance variability (Fig. S3). To gain further insight, we performed an exploratory analysis of the parameters of the learned HMO model, evaluating each parameter both for how sensitively it was tuned and how diverse it was between model mixture components. Two classes of model parameters were especially sensitive and diverse (*SI Text* and Figs. S10 and S11): (*i*) filter statistics, including filter mean and spread, and (*ii*) the exponent trading off between max-pooling and average-pooling (16). This observation hints at a computationally rigorous explanation for experimentally
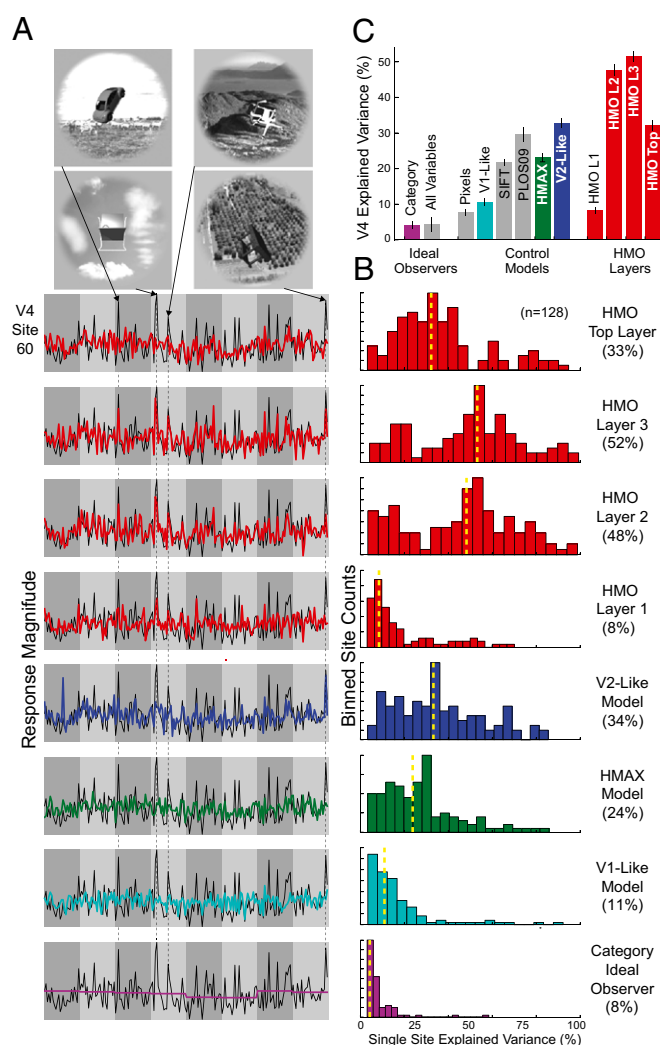

Fig. 5. V4 neural predictions. (*A*) Actual vs. predicted response magnitudes for a typical V4 site. V4 sites are highly visually driven, but unlike IT sites show very little categorical preference, manifesting in more abrupt changes in the image-by-image plots shown here. Red highlight indicates the best-matching model (*viz.*, HMO layer 3). (*B*) Distributions of explained variances percentage for each model, over the population of all measured V4 sites (*n* = 128). (*C*) Comparison of V4 neural explained variance percentage for various models. Conventions follow those used in Fig. 3.

Yamins et al.

observed heterogeneities in higher ventral cortex areas (13, 32), but much work remains to be done to confirm such a hypothesis.

Top-Down Approach to Understanding Cortical Circuits. A common assumption in visual neuroscience is that understanding the tuning curves of neurons in lower cortical areas will be a necessary precursor to explaining higher visual cortex. For example, significant work has gone into assessing the extent to which V4 neurons can be understood as a curvature-selective shape representation (27). Our results indicate that it is useful to complement this bottom-up approach with a top-down perspective characterizing IT as the product of an evolutionary/developmental process that selected for high performance on recognition on tasks like those used in our optimization. V4 may in turn be characterized as having been selected precisely to support the downstream computation in IT. This type of explanation is qualitatively different from more traditional approaches that seek explicit descriptions of neural responses in terms of particular geometrical primitives. However, our results show functionally relevant constraints can be used to obtain quantitatively predictive models even when such explicit bottom-up primitives have not been identified.

Going forward, we will bridge these bottom-up and top-down explanations by building links to lower and intermediate visual cortex, especially in V1 and V2. We will also explore recent high-performing computer vision systems with architectures inspired by the ventral stream (36). Our results show that behaviorally driven computational approaches have an important role in understanding the details of visual processing (37) and suggest that the overall approach may be applicable to other cortical areas and task domains.

Materials and Methods

Array Electrophysiology. Neural data were collected in two awake behaving rhesus macaques (*Macaca mulatta*, 7 and 9 kg) using parallel multielectrode array electrophysiology systems (Cerebus System; BlackRock Microsystems). All procedures were done in accordance with National Institutes of Health guidelines and approved by the Massachusetts Institute of Technology (MIT) Committee on Animal Care. 296 neural sites (168 in IT and 128 in V4) were selected as being visually driven. Fixating animals were presented with testing images for 100 ms, and scalar firing rates were obtained from spike trains by averaging spike counts in the period 70–170 ms after stimulus presentation. See *SI Text* for additional details.

Neural Predictivity Metric. For each IT neural site, we used linear regression to identify a linear weighting of model output units (from the top or intermediate layers) that is most predictive of that site's actual output on a fixed set of sample images (10, 26, 27). Using this "synthetic neuron," we then produced per-image response predictions on novel images not used in the regression training and compared them to the actual neural site's output for those images (Figs. 3*A* and 5*A*). We computed the goodness-of-fit r^2 value, normalized by the neural site's trial-by-trial variability, to obtain the explained variance percentage for that site. The overall area predictivity of a model is the median explained variance over all measured sites in that area (Figs. 3 *B* and *C* and 5 *B* and *C* and see *SI Text*).

1. DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. *Trends Cogn Sci* 11(8):333–341.
2. Grill-Spector K, Kourtzi Z, Kanwisher N (2001) The lateral occipital complex and its role in object recognition. *Vision Res* 41(10-11):1409–1422.
3. Malach R, Levy I, Hasson U (2002) The topography of high-order human object areas. *Trends Cogn Sci* 6(4):176–184.
4. Kriegeskorte N, et al. (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60(6):1126–1141.
5. Tanaka K (1996) Inferotemporal cortex and object vision. *Annu Rev Neurosci* 19:109–139.
6. Logothetis NK, Sheinberg DL (1996) Visual object recognition. *Annu Rev Neurosci* 19:577–621.
7. Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1(1):1–47.
8. Vogels R, Orban GA (1994) Activity of inferior temporal neurons during orientation discrimination with successively presented gratings. *J Neurophysiol* 71(4):1428–1451.
9. DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? *Neuron* 73(3):415–434.
10. Carandini M, et al. (2005) Do we know what the early visual system does? *J Neurosci* 25(46):10577–10597.
11. Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. *Science* 310(5749):863–866.
12. Rust NC, DiCarlo JJ (2010) Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J Neurosci* 30(39):12978–12995.
13. Freiwald WA, Tsao DY (2010) Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* 330(6005):845–851.
14. Connor CE, Brincat SL, Pasupathy A (2007) Transformation of shape information in the ventral pathway. *Curr Opin Neurobiol* 17(2):140–147.
15. Fukushima K (1980) Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern* 36(4):193–202.
16. Riesenhuber M, Poggio T (2000) Models of object recognition. *Nat Neurosci* 3(Suppl):1199–1204.
17. Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci USA* 104(15):6424–6429.
18. Lecun Y, Huang F-J, Bottou L (2004) Learning methods for generic object recognition with invariance to pose and lighting. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, Washington, DC), Vol 2, pp 97–104.
19. Bengio Y (2009) Learning deep architectures for AI. *Foundations and Trends in Machine Learning* (Now Publishers, Hanover, MA), Vol 2.
20. Pinto N, Doukhan D, DiCarlo JJ, Cox DD (2009) A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLOS Comput Biol* 5(11):e1000579.
21. Kiani R, Esteky H, Mirpour K, Tanaka K (2007) Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J Neurophysiol* 97(6):4296–4309.
22. Rust NC, Mante V, Simoncelli EP, Movshon JA (2006) How MT cells analyze the motion of visual patterns. *Nat Neurosci* 9(11):1421–1431.
23. Gallant JL, Connor CE, Rakshit S, Lewis JW, Van Essen DC (1996) Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *J Neurophysiol* 76(4):2718–2739.
24. Cadieu C, et al. (2013) The neural representation benchmark and its evaluation on brain and machine. International Conference on Learning Representations 2013. arXiv:1301.3530.
25. Pinto N, Cox DD, DiCarlo JJ (2008) Why is real-world visual object recognition hard? *PLOS Comput Biol* 4(1):e27.
26. Cadieu C, et al. (2007) A model of V4 shape selectivity and invariance. *J Neurophysiol* 98(3):1733–1750.
27. Sharpee TO, Kouh M, Reynolds JH (2013) Trade-off between curvature tuning and position invariance in visual area V4. *Proc Natl Acad Sci USA* 110(28):11618–11623.
28. Mutch J, Lowe DG (2008) Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision* 80(1):45–57.
29. Brincat SL, Connor CE (2004) Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat Neurosci* 7(8):880–886.
30. Lowe DG (2004) *Distinctive Image Features from Scale-Invariant Keypoints* (IJCV).
31. Freeman J, Simoncelli EP (2011) Metamers of the ventral stream. *Nat Neurosci* 14(9):1195–1201.
32. Downing PE, Chan AW, Peelen MV, Dodds CM, Kanwisher N (2006) Domain specificity in visual cortex. *Cereb Cortex* 16(10):1453–1461.
33. Schapire RE (1999) *Theoretical Views of Boosting and Applications*, Lecture Notes in Computer Science (Springer, Berlin), Vol 1720, pp 13–25.
34. Geisler WS (2003) Ideal observer analysis. *The Visual Neurosciences*, eds Chalupa L, Werner J (MIT Press, Boston), pp 825–837.
35. Pasupathy A, Connor CE (2002) Population coding of shape in area V4. *Nat Neurosci* 5(12):1332–1338.
36. Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25:1097–1105.
37. Marr D, Poggio T, Ullman S (2010) *Vision. A Computational Investigation Into the Human Representation and Processing of Visual Information* (MIT Press, Cambridge, MA).

Yamins et al.