Article

Current Biology

A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition

Highlights

- Voxel decomposition infers dominant response components in the ventral visual cortex
- Top components replicate neural selectivities for faces, bodies, scenes, and words
- A novel food-selective component is discovered in the ventral visual cortex
- Alternative accounts based on color, shape, or texture fail to account for food selectivity

Authors

Meenakshi Khosla,
N. Apurva Ratan Murty,
Nancy Kanwisher

Correspondence

mkhosla@mit.edu

In brief

Using data-driven voxel decomposition on a large-scale naturalistic dataset, Khosla et al. discover a novel selectivity to visual images of food in the ventral visual cortex. This hypothesis-free approach also rediscovers neural selectivities to faces, bodies, scenes, and words, revealing their prominent role in the organization of the visual cortex.

Khosla et al., 2022, Current Biology *32*, 1–13
October 10, 2022
https://doi.org/10.1016/j.cub.2022.08.009

CellPress

Current Biology

CellPress

Article

A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition

Meenakshi Khosla,[1,2,*] N. Apurva Ratan Murty,[1] and Nancy Kanwisher[1]
[1]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Lead contact
*Correspondence: mkhosla@mit.edu
https://doi.org/10.1016/j.cub.2022.08.009

SUMMARY

Prior work has identified cortical regions selectively responsive to specific categories of visual stimuli. However, this hypothesis-driven work cannot reveal how prominent these category selectivities are in the overall functional organization of the visual cortex, or what others might exist that scientists have not thought to look for. Furthermore, standard voxel-wise tests cannot detect distinct neural selectivities that coexist within voxels. To overcome these limitations, we used data-driven voxel decomposition methods to identify the main components underlying fMRI responses to thousands of complex photographic images. Our hypothesis-neutral analysis rediscovered components selective for faces, places, bodies, and words, validating our method and showing that these selectivities are dominant features of the ventral visual pathway. The analysis also revealed an unexpected component with a distinct anatomical distribution that responded highly selectively to images of food. Alternative accounts based on low- to mid-level visual features, such as color, shape, or texture, failed to account for the food selectivity of this component. High-throughput testing and control experiments with matched stimuli on a highly accurate computational model of this component confirm its selectivity for food. We registered our methods and hypotheses before replicating them on held-out participants and in a novel dataset. These findings demonstrate the power of data-driven methods and show that the dominant neural responses of the ventral visual pathway include not only selectivities for faces, scenes, bodies, and words but also the visually heterogeneous category of food, thus constraining accounts of when and why functional specialization arises in the cortex.

INTRODUCTION

The last few decades of research in human cognitive neuroscience have revealed the functional organization of the cortex in rich detail. This organization features a set of regions that are selectively engaged in single mental processes, from perceiving faces or scenes or music, to understanding the meaning of a sentence, to inferring the content of another person's thoughts. Why do our brains have these particular specializations and apparently not others? To answer this question, we need a more complete inventory of human cortical specializations, one that reflects not just the idiosyncratic hypotheses scientists have already thought to test but also the actual functional organization of the cortex itself. Here, we tackle this question for the ventral visual pathway by searching in a hypothesis-neutral fashion for the dominant neural response profiles in this region in a large, recently released public dataset of fMRI responses to thousands of natural images in each of 8 participants.[1]

Extensive evidence[2–7] from neurological patients, fMRI, and intracranial recording and stimulation has demonstrated that the ventral visual pathway contains distinct regions causally engaged in the perception of faces, scenes, bodies, and words. But are these categories the main ones, or might others exist that have not yet been found? The current evidence does not answer this question for several reasons. First, prior research on the ventral pathway has tested a relatively small number of stimulus categories, which may not have subtended the relevant part of stimulus space preferred by some neural populations. Second, this work has proceeded in a largely hypothesis-driven fashion and thus may have missed neural populations with response profiles scientists have not thought to test. Third, prior research based on voxel-wise contrasts is not well suited to discovering neural populations whose high selectivity is masked because the fMRI signal averages their responses with the responses of other neural populations cohabiting the same voxels.[8]

Here, we overcome these three limitations by analyzing fMRI responses to the very broad and large set of natural stimuli in the Natural Scenes Dataset (NSD[1]) with a data-driven analysis method that can de-mix the underlying responses from neural populations that are spatially intermingled within individual fMRI voxels. Specifically, we factorized the matrix of response magnitudes of each voxel to each stimulus into a set of components, which we hypothesize correspond to distinct neural populations. Each component is described by a response profile across stimuli and a weight matrix indicating how strongly that component contributes to each voxel's response (Figure 1).

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

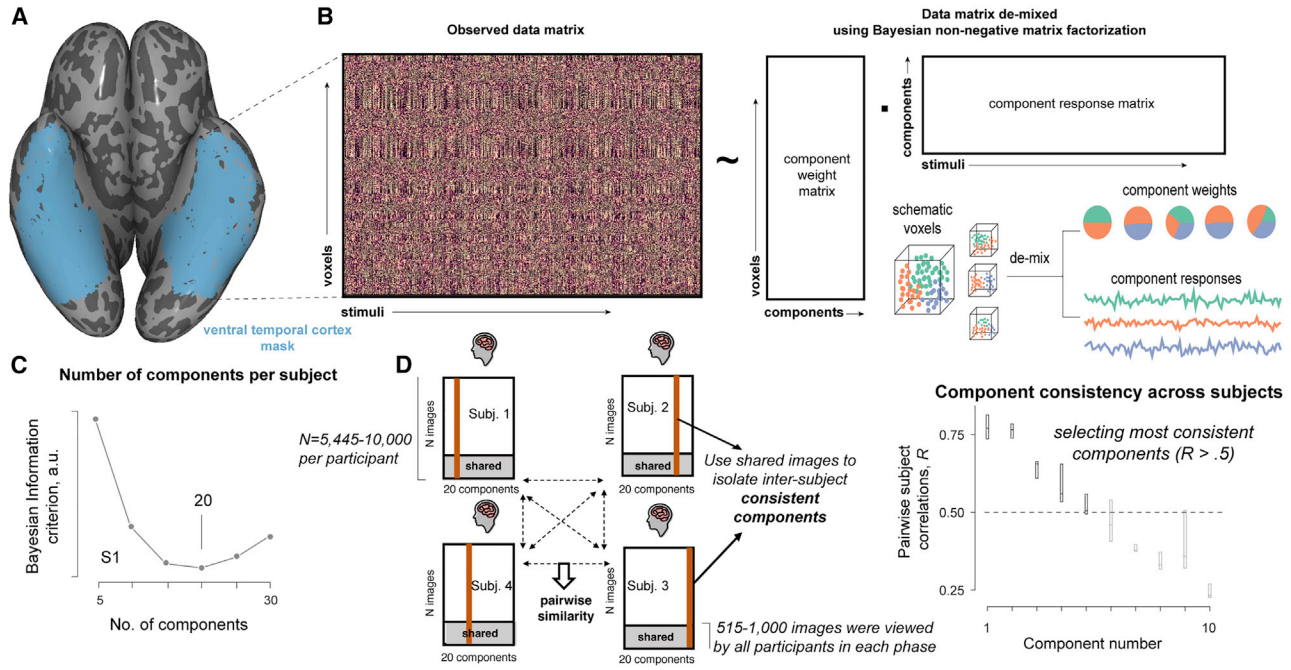

Figure 1. Outline of the data-driven component modeling approach

(A) The large swath of ventral visual cortex included in our analyses for one example subject.

(B) Illustration of the data-driven voxel decomposition approach. Bayesian non-negative matrix factorization was used to decompose the observed ventral visual stream data matrix of each participant as a product of two lower-dimensional matrices: (1) a response profile matrix that characterizes the response of each component to all 5,445–10,000 stimuli viewed by each participant and (2) a component by voxel weight matrix that expresses the contribution of each component to each of the ~6,500–9,600 voxels per participant.

(C and D) (C) The Bayesian information criterion (BIC) as we vary the number of components in one participant. The optimal number of components was chosen as the minimum BIC. Other subjects had a similar trend. Components present in all subjects were isolated by measuring pairwise inter-subject correlations of component response profiles, as illustrated in (D) for phase I participants (right). Gray-shaded region shows the proportion of stimuli viewed by all 4 participants in each phase (515 images in phase 1 and 1,000 in phase 2). The top 5 components based on this metric all had mean and median pairwise inter-subject consistency > 0.5 (D, right).

This analysis method enables us to discover the main components that explain neural responses in the ventral visual pathway, potentially including new selectivities not described previously.

RESULTS

We applied hypothesis-neutral Bayesian non-negative matrix factorization (NMF)[9] methods to the NSD[1] to identify the dominant neural populations in the human ventral visual pathway. Importantly, the algorithm does not have any information about the images or the spatial location of voxels. Instead, it infers the response profiles and anatomical distribution of distinctive neural populations solely from the unlabeled voxel response matrix. This method is thus a powerful way to both validate known selectivities and discover new ones. Our approach is similar to that of Norman-Haignere et al.,[8] except that we use NMF instead of independent components analysis (see STAR Methods for rationale). In phase I of this project, we analyzed data for four of the eight available NSD participants and presented this work at the Vision Sciences Society meeting.[10] We then pre-registered our analyses on the Open Science Framework (https://osf.io/n47qf) and confirmed our hypotheses on the four held-out NSD participants (phase 2). We report results for both groups analyzed separately as well as for each individual participant in supplemental information.

Our general procedure is illustrated in Figure 1. We first applied the NMF algorithm on each subject's data separately (Figures 1A and 1B) to identify subject-specific components (phase 1 subjects viewed 10,000 images, phase 2 subjects viewed 5,445 images). Bayesian information criterion (BIC) applied to Bayesian NMF yielded ~20 components in each subject (Figure 1C). Next, we used overlapping images viewed by all subjects (in phase 1 and 2 separately) to identify and rank the consistent components across subjects using a pairwise inter-subject consistency metric. This method identified 5 consistent components across subjects with median pairwise consistency > 0.5. The 5 components derived from phase 1 participants collectively accounted for ~50% of the replicable variance in the reliable voxels in phase 2 subjects (reliability threshold > 0.3, 46% ventral stream voxels) and were highly correlated with the top five components identified independently in phase 2 subjects (Figure S1), confirming reproducibility of this 5-component structure.

Characterizing the function of the top components

We first qualitatively examined the response profiles of the top 5 components. For each component, we sorted stimuli by their

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

CellPress

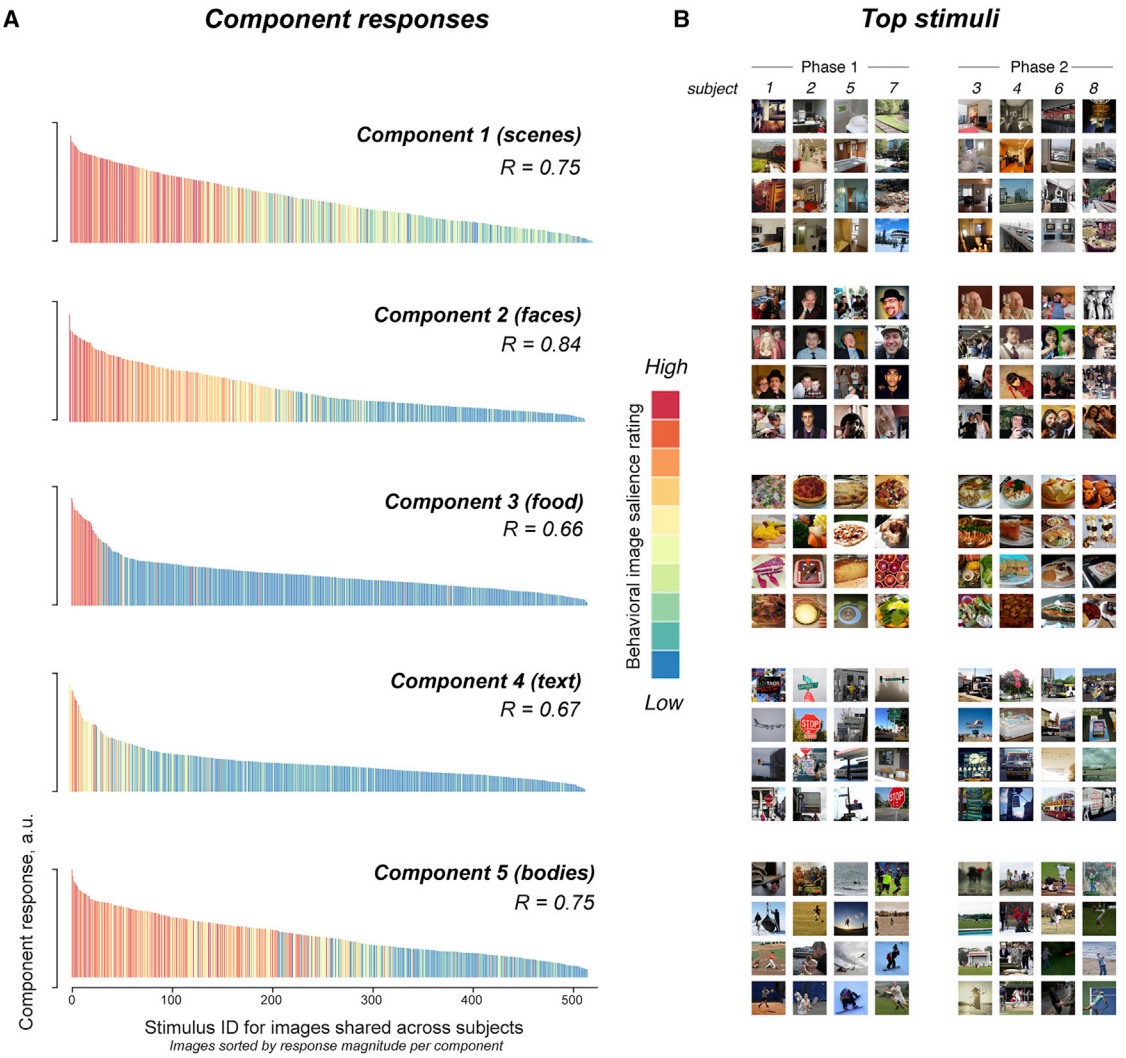

Figure 2. Response profile and preferred stimuli for the top five components

(A) Response profile for each of the top 5 components (with highest inter-subject consistency) across the 515 images seen by all participants. These components were derived separately within each of the 8 participants individually (see Figure S2 for corresponding data from each participant) but are shown here averaged across all 8 participants. Each bar is an image, and the colors indicate the behavioral salience rating for the preferred category (e.g., the salience of faces for Component 2).

(B) Top 4 images producing the strongest response in each component in each phase 1 (left) and phase 2 (right) participant.

See Video S1 for the top 25 images for each component in each subject.

response magnitude in this component and inspected the top 25 images of each component for each participant. These images (top 4 shown in Figure 2 and all the top 25 shown for each participant in Video S1) revealed a distinctive and familiar selectivity pattern for four of the top five components (Figure 2). The images that produced the highest responses in components 1, 2, 4, and 5 were, respectively, scenes, faces, text (including words and symbol strings), and bodies (either full bodies or body parts). To validate this apparent preferred category of each component, we collected ratings for each of these preferred categories in a behavioral experiment where participants were asked to rate the salience of each of these categories in each of the images viewed by all NSD participants (STAR Methods). Salience ratings for the scenes, faces, text, and bodies were strongly correlated with the response of components 1, 2, 4, and 5 (respectively) across images. These findings are consistent with a large prior literature on selectivities for these categories in the ventral visual pathway[11] (and their anatomical location, discussed below), so we considered these results as positive controls on our method and did not interrogate these response profiles further.

A novel component selectively responsive to food

Component 3, however, was unexpected. This component, the third-most-consistent component across participants in the separate analyses of both phase 1 and phase 2 participants, appeared to respond in a highly selective fashion to images of food. This food selectivity is evident both in the correlation of the component's response profile with rated salience of food (Figure 2A)

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

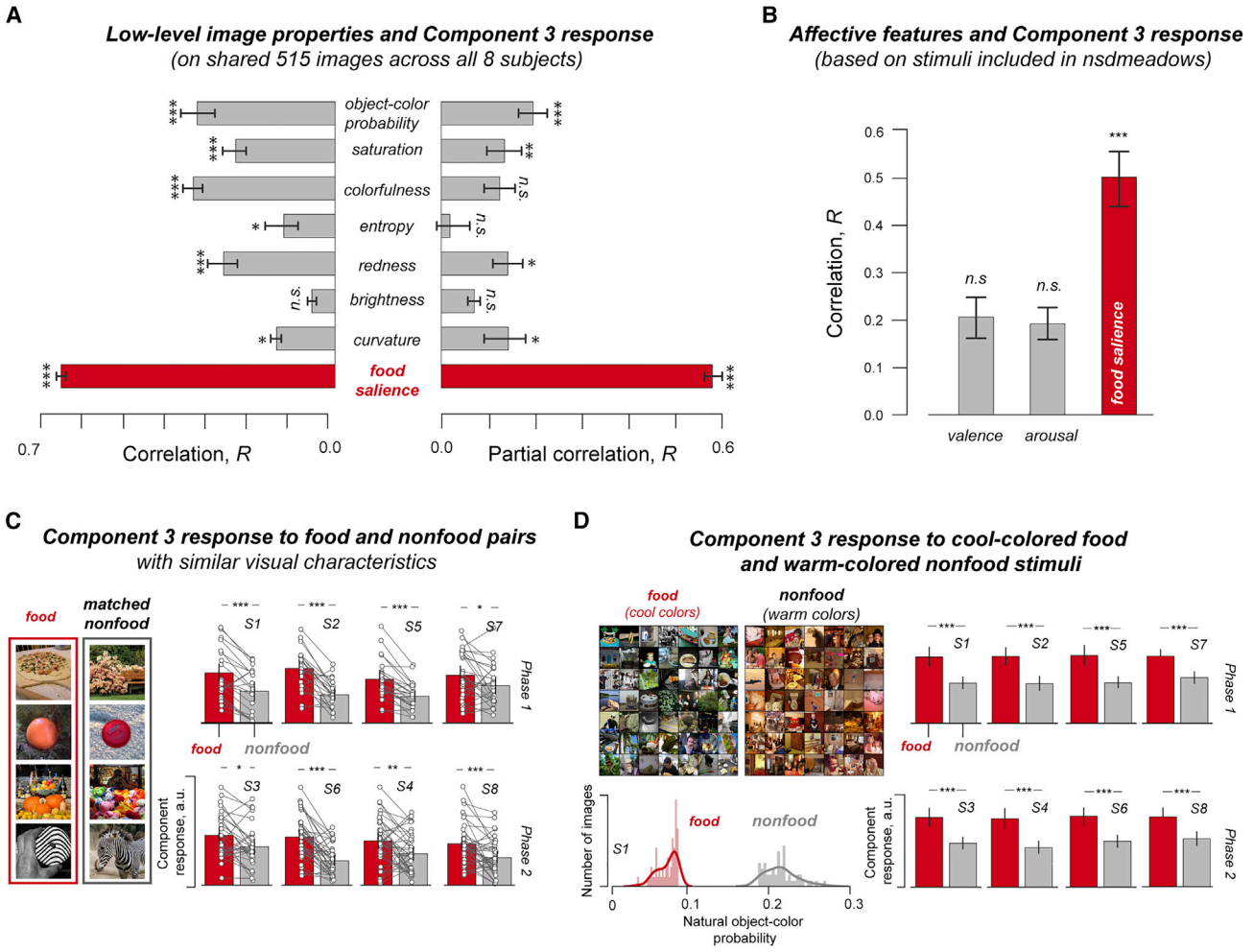

Figure 3. Alternate accounts do not explain the food selectivity of Component 3

(A) Left: the correlation across stimuli between the magnitude of the Component 3 response averaged across the 8 participants and various image-computable feature dimensions and rated food salience (STAR Methods). Right: same but now with food salience partialled out for the image-computable measures, and with all other measures partialled out for the food salience measure.

(B) The correlation between the magnitude of the Component 3 response and valence and arousal ratings across the subset of 100 stimuli for which these ratings were available in the original NSD study, called "*nsdmeadows*" stimuli.[1]

(C) Responses of Component 3 in each participant to food and non-food stimuli selected in pairs of images (one food and one non-food) that produce similar activations in the last convolutional layer ("*conv5*") of an AlexNet architecture pre-trained on ImageNet. See also Video S3.

(D) Response of Component 3 in each participant to sets of stimuli chosen such that the food images were very low, and the non-food images were very high on the object-color probability measure.[12]

See also Figures S3 and S6.

and in the images that produced the highest response in individual subjects (Figure 2B). Although most of the top-ranked images are of prepared food (e.g., a slice of pizza), unprepared food (e.g., a broccoli, carrot, banana, etc.) also produced strong responses in this component (Video S1). But inspection of those top images also suggests several potential alternative accounts for this component's responses. For example, the top images for this component also seem to share certain low-level and mid-level visual features, including warmer and more saturated colors, higher curvature, and a complex spatial structure with rich texture. To address these potential alternative accounts of food selectivity, we first estimated several image-computable metrics of color, curvature, and texture (STAR Methods). The component response was more strongly correlated to the behavioral salience ratings for food than any of these other visual feature metrics (Figure 3A). However, some of the visual properties were also significantly correlated with the component response, particularly the object-color probability metric[12] (the probability of a hue being a natural object, which reflects the warm-cool color continuum).

But how much unique variance do each of these variables explain of the component's responses? Figure 3A (right) demonstrates that the rated salience of food remains highly correlated with the response profile of Component 3 (R = 0.58, p = 6e–47), even after partialling out all the visual features that appear to be most confounded with food. Some of these visual properties on

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

CellPress

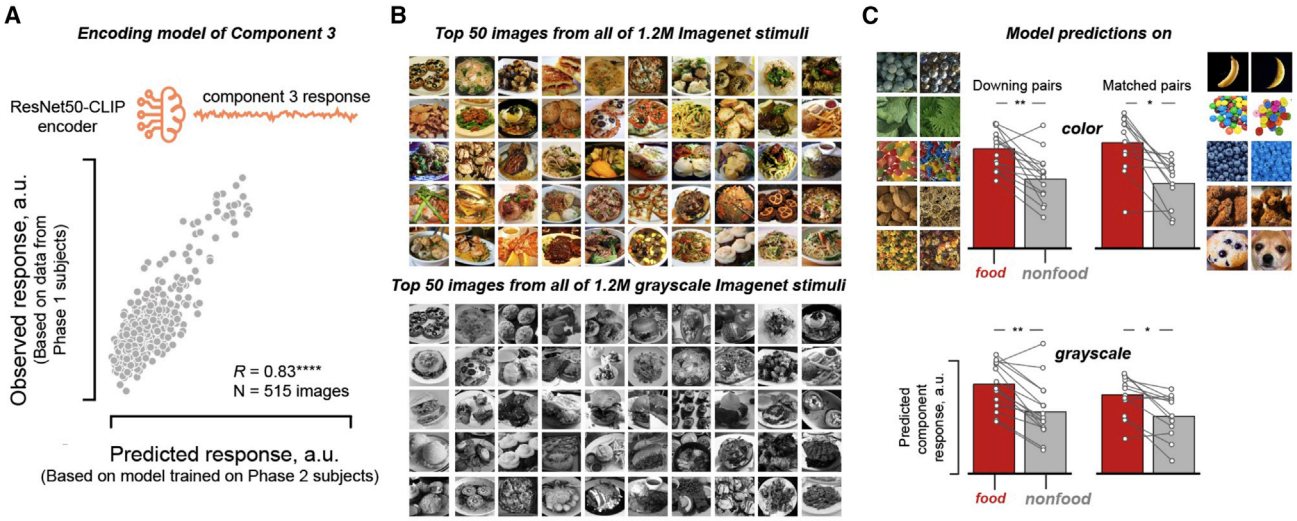

Figure 4. A CNN-based encoding model of Component 3 response enables tests on images beyond those in the NSD

(A) We used a ResNet50-CLIP encoder to predict the Component 3 response. The x axis shows the predicted component response (based on the model trained on phase 2 subjects), and the y axis shows the observed Component 3 response (from phase 1 subjects). Each dot is an image (total n = 515 shared images across all 8 subjects) that the model did not encounter in the model fitting procedure (cross-validated on both images and subjects).

(B) Top 50 stimuli predicted by the encoding model to have the highest response across all 1.2 M colored (top) and grayscale (bottom) images from the ImageNet dataset. All images are of food. See also Video S4.

(C) Model prediction on colored (top) and grayscale (bottom) versions of the Downing pairs and our Matched pairs of food and non-food images. The bars indicate the mean response, and each connected line indicates a particular matched food-non-food pair.

their own account for a significant, though much smaller, part (e.g., for object-color probability, R = 0.16, p = 8e–6) of the component's response once food salience is partialled out, although the unique variance explained by food salience ratings is significantly greater than the variance explained by any of these visual properties with the effect of food salience removed (all $p < 0.00001$; Figure S3). Figure 3B further shows that the response of Component 3 was not significantly correlated with behavioral ratings of either valence or arousal provided for a subset of the stimuli in the original NSD study. Further, the food selectivity of this component persists strikingly across the large set of 5,445–10,000 images viewed by each participant (data not shown).

These results indicate that the visual features most obviously related to Component 3 cannot alone explain the response of Component 3. However, it remains possible that this component selectively responds to a conjunction of multiple lower-level features (like reddish, round objects). We therefore performed three further analyses. First, we identified pairs of images (one food and one non-food) from the 5,445–10,000 image sets viewed by each participant that produce similar activations in the last convolutional layer ("*conv5*") of a pre-trained AlexNet (STAR Methods). These food and non-food pairs are visually very similar, with matching features in similar spatial locations (Figure 3C). Yet food images still produced a significantly higher response than their matched non-food images in each participant (paired t tests, all $p < 0.01$). Second, we identified food images that ranked low on an object-color probability measure (Rosenthal et al.,[12] i.e., "cool" colored food) and non-food images that ranked high on the same scale (i.e., "warm" non-food images). The component response remained significantly higher to the food images than the non-food images in every subject (all $p < 0.001$; Figure 3D), suggesting that the component's food selectivity overrides any sensitivity to object-color probability (see also Figure S3). Third, we selected subsets of food and non-food stimuli that maximally span the embedding space of different layers of an ImageNet-trained AlexNet model,[13] such that the sampled images within each set are substantially dissimilar among themselves, and the selected subsets are diverse on perceptually relevant image properties (Video S3; described further in the STAR Methods). As a result, a linear classifier trained to discriminate between these food and non-food images using the features of the corresponding layer performs at chance (never exceeding 53%). And yet the food images still produce a significantly higher response than the non-food images in Component 3, even across these highly heterogeneous food and non-food subsets (Video S3), showing that the food preference holds broadly and is not limited to specific kinds of food images.

Our analyses thus far focused only on the stimuli included in the NSD. Although the NSD includes a large number of images (n = 56,720 across 8 subjects with full repetitions each), they span a small subset of the space of all possible images. To address this limitation, we built a deep convolutional neural network (CNN)-based encoding model to predict the response of Component 3 (see Ratan Murty et al.[14]). Our CLIP-ResNet50[15]-based encoding model was highly accurate at predicting the response to images not encountered in the model training procedure (correlation between the cross-validated predicted and observed responses = 0.83, $p < 0.00001$; Figure 4A). The success of this computational model in mimicking component responses and its image-computable nature thus allow us

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

to use the model-predicted responses as substitutes for their actual measured neural counterparts on much larger stimulus sets, well beyond NSD. Would the food selectivity of the component hold even when tested on a much larger battery of stimuli? To find out, we obtained predictions for the Component 3 response to all 1.2 million stimuli from the ImageNet dataset (P. Downing and N. Kanwisher, 1999, Cogn. Neurosci. Soc., poster).[16] All the top 1,000 stimuli predicted to activate this component (from ~1.2 million possible images) contained food (see Figure 4B, top, for the top 50 images and Video S4 for the top 200 images), while none of the bottom 1,000 stimuli were found to contain food. This high-throughput screening procedure on CNNs validates the observed food selectivity of Component 3. These analyses, however, are subject to a few caveats. Models can only substitute for neural data insofar as the model is accurate, and because not all variance in component responses is accounted for by the model, conclusions based on model predictions are not error-free and will ultimately need to be validated with actual neural data. Further, even when the computational model can capture all meaningful variance in neural responses, it does not necessarily follow that the model uses the same mechanisms as the brain to arrive at the predicted response. In this context, however, where the goal is to use models as virtual stand-ins for experimental data, the mechanistic similarity is secondary to prediction performance as an index of the utility of the model.

Downing and Kanwisher (P. Downing and N. Kanwisher, 1999, Cogn. Neurosci. Soc., poster) had previously tested and rejected the food selectivity hypothesis when they failed to find higher responses to food textures compared with visually similar non-food textures. When these stimuli were tested on the CNN-based model of the Component 3, it predicted a significantly higher response to food than the non-food matched textures (Figure 4C, top left; paired t test, $t(14) = 5.21$, $p = 1.53 \times 10^{-5}$). Next, we handpicked a new set of food and non-food images that look very similar (examples shown in Figure 4C, right). Here, too, our computational model predicted a significantly higher response to food than matching non-food images (Figure 4C, top right; paired t test, $t(11) = 3.34$, $p = 0.003$). Would food selectivity hold even when tested on grayscale images? We tested this in 3 different ways. First, we again obtained predictions for the entire 1.2 million ImageNet stimuli, but this time on grayscale versions of the same images. The correlation between the predicted response to the color versus the grayscale version of each image was very high ($n = 1{,}281{,}167$ images, Pearson's $R = 0.98$, $p < 0.00001$). Critically, the top 1,000 images predicted to have the highest response, even from the grayscale set, were all of food (see Figure 4B, bottom, for the top 50 images). Second, our computational model predicted a significantly higher response to food than the non-food matched textures from the black and white versions of the Downing image pairs (Figure 4C, bottom left; paired t test, $t(14) = 2.93$, $p = 0.007$). Third, our computational model also predicted a significantly higher response to black-and-white food than non-food in our handpicked matched images (Figure 4C, bottom right; paired t test, $t(11) = 2.44$, $p = 0.023$). Together, these computational modeling results complement our previous analyses by confirming the observed food selectivity of Component 3 across a larger number of images, for stringent control images, and by showing that the selectivity of this component for food over non-food persists even for grayscale images.

The observed components are not artifacts of the stimulus set composition

Might the observed components reflect the composition of the stimulus set rather than a property of the brain itself? Of course, experimental approaches, whether hypothesis-neutral or hypothesis-driven, cannot reveal selectivities for stimulus classes that are not included in the stimulus set, and there is likely to be some effect of the relative proportion of different stimulus types in the set. However, it seems unlikely that the category selectivities found (for faces, scenes, bodies, text, and food) reflect an over-representation of these categories in the stimulus set compared with human experience, given that most humans spend at least an hour per day engaged in activities where these visual stimuli feature prominently.[17] In addition, several further analyses show that the food selectivity of Component 3 is not an artifact of the composition of the stimuli. First, the food-selective component emerges separately in each of the eight participants, despite the fact that they saw mostly different stimuli, and it can also be identified in the completely separate BOLD5000 dataset[18,19] (Figure 5). And, conversely, no food-selective component is found when the same analyses are applied to responses to the same images in retinotopic cortex, dorsal, and lateral visual streams or in early layers of a CNN. Thus, the use of the NSD stimulus set on its own is neither necessary nor sufficient to find food selectivity.

Finally, to test whether any stimulus category that represents a sizable proportion of the stimuli will result in a component selectively responsive to that category, we performed one further analysis (Figures 5D and 5E). First, we selected a subset of the stimuli that contained equal numbers of exemplars in each of 9 categories, including faces (as a positive control), food, and 7 other perceptually homogeneous categories (airplanes, clocks, horses, elephants, giraffes, trucks, motorcycles). Repeating the NMF analysis on these data revealed one component selectively responsive to faces and another to food, and no components as highly selectively responsive to any of the other categories, even though the food images were drawn to be one of the least homogeneous in this set. Thus, ample representation of a category, even a perceptually homogeneous category such as airplanes or clocks, in the stimulus set is not sufficient for a component selective to that category to emerge and cannot account for the food selectivity of Component 3.

Demixing reveals stronger selectivity for components than voxels

Why was food selectivity not observed before, particularly in previous hypothesis-driven investigations (P. Downing and N. Kanwisher, 1999, Cogn. Neurosci. Soc., poster)?[20] We speculated that the spatial overlap of food-selective neural populations with other selectivities dilutes food selectivity in individual voxels, which our demixing procedure is able to uncover. We tested this idea by measuring the selectivity of the demixed Component 3 and of the average response across the top 1% of voxels, with the highest weight on Component 3. Food selectivity was significantly higher for the demixed Component 3 (mean R across subjects = 0.53, $p < 0.00001$) than in the top voxels (mean $R = 0.41$, $p < 0.00001$) in each participant ($t(8) = 10.7$, $p < 0.0001$; Figure 6). For comparison, face selectivity of

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

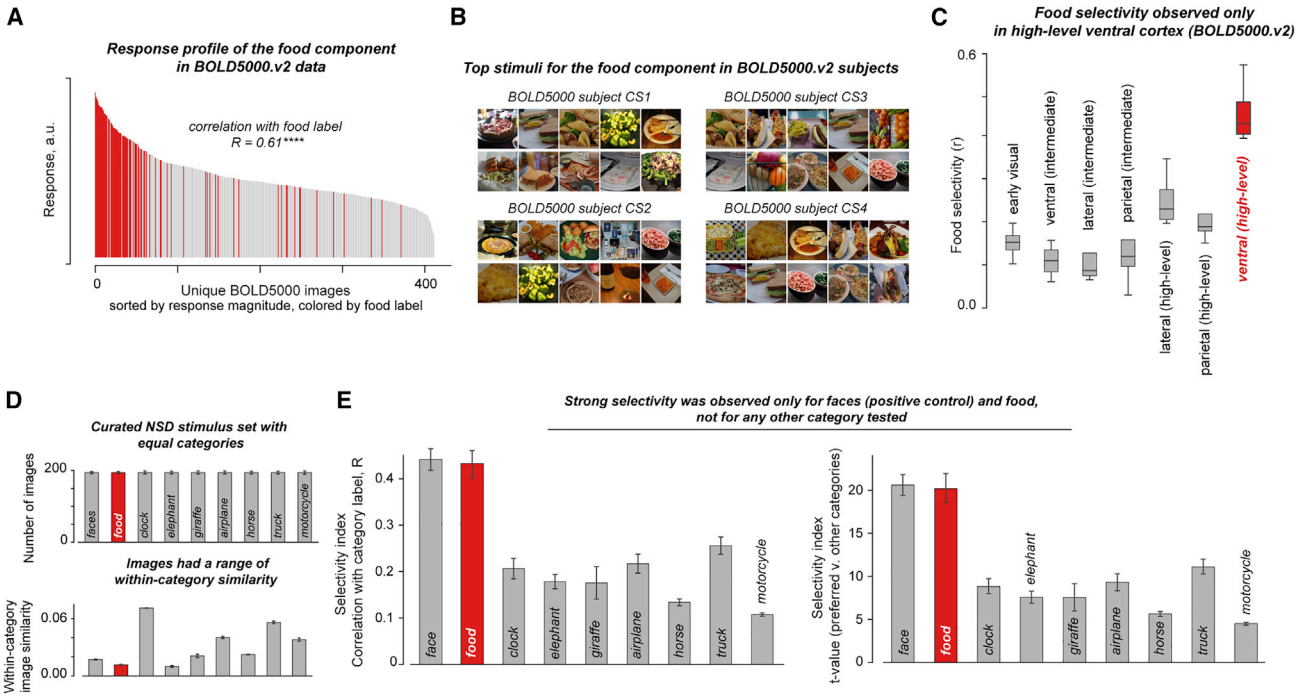

Figure 5. Observed food component is not an artifact of the NSD stimulus composition

We first tested the dependence of stimulus composition, if any, on the completely independent BOLD5000.v2 fMRI dataset.

(A) Response profile for the food component identified in the BOLD5000 data (from images shared with NSD) on images unique to BOLD5000.v2. The x axis shows the stimuli, and the y axis shows the inferred response magnitude for the Component 3 for images unique to BOLD5000. The bars in red are images that were labeled as food, and the bars in gray are images labeled non-food in the (imperfect) annotations provided with MS-COCO.

(B) Top 10 images for each of the four subjects in the BOLD5000v2 dataset.

(C) Boxplots showing the food selectivity distribution across BOLD5000 subjects (y axis) for the components inferred from different cortical regions (x axis). Food selectivity was observed only in ventral visual cortex, not in other regions.

(D) Next, we performed the NMF decomposition on a curated subset of the NSD with 9 stimulus categories, each with an equal number of images within each subject (top) with high within-category visual similarity for the non-food categories (bottom).

(E) NMF decomposition on this curated stimulus set revealed components with strong selectivity only for faces (positive control) and food, not for any other category. The y axis shows the highest selectivity obtained for each of the 9 categories (among all components) based on category labels from MS-COCO using two different metrics (left, correlation with a binary category label vector indicating whether the category was present/absent in the image; and right, t value comparing the mean response to stimuli from that category versus all other stimuli).

See STAR Methods for details.

Component 2 was as strong in the top voxels as within the inferred component response ($t(8) = 0.85$, $p = 0.42$; Figure 6). These results indicate that the neural populations selective for food are likely more mixed with other neural populations within voxels than the face-selective neural populations, explaining why strong food selectivity has not been found previously with standard analysis methods.

Anatomical distribution of components

We next characterized the anatomical distribution of each component by projecting its voxel weights back into anatomical coordinates within each participant individually. For known selectivities, the component anatomies exhibited clear agreement with the corresponding regions identified with an independent functional localizer: the face-selective component produced the highest voxel weights in the fusiform face area (FFA) and other known face-selective sites such as aTL faces (anterior temporal lobe faces) and mTL faces (mid temporal lobe faces), the text-selective component was concentrated within the visual word form area (VWFA), the scene-selective component in the parahippocampal place area (PPA), and the body-selective component in parts of the fusiform body area (FBA) and the extrastriate body area (EBA), as shown qualitatively in Figure 7 and Video S2. Quantitatively, the voxel weight maps demonstrated high correlations with the t statistics pertaining to the relevant domain from the functional localizer experiment (Figure S4).

The weight maps for the novel food-selective component appeared patchy across the cortex, with considerable variability across participants. To get an impression of the component anatomy, we first registered each participant's voxel weight map to a common MNI space. The subject-averaged voxel weight map for Component 3 (Figure 7C) shows its distinct anatomy, which is largely concentrated in two clusters, one medial and one lateral to the FFA. We also emulated a contrast-based experiment wherein we selected a subset of food and non-food images from the shared set of 515 NSD images viewed by all participants (STAR Methods). We computed the food contrast for each participant individually, using the conventional t statistic comparing responses of all stimulus-driven voxels to food versus non-food stimuli. The subject-averaged maps from this contrast-based experiment highlight that significant

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

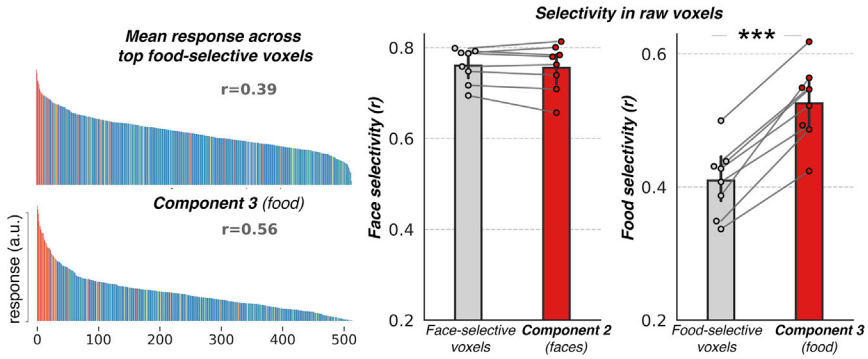

Figure 6. Stronger selectivity for components than average response across voxels

Left: average response profile of the top 1% voxels with the highest weights on Component 3 (top) and the response profile of Component 3 (bottom) for one subject. Food selectivity of the respective response profiles is reported at the top of each subplot. Right: face selectivity of the mean responses across top 1% voxels with highest weights on Component 2 (the face-selective component) and the corresponding selectivity of Component 2. Rightmost plot shows the food selectivity of the mean response across top 1% voxels with highest weights on Component 3 (the food-selective component) and the corresponding selectivity of Component 3. Each dot in the swarm plot is an individual subject. Here, selectivity is computed as the correlation between responses and the salience ratings for the preferred category over the 515 stimuli shared across all 8 participants.

differences are largely restricted to small clusters in the ventral visual stream (Figure 7D) in roughly similar anatomical locations as that of Component 3.

To quantify the inter-subject variability in the anatomy of Component 3 and compare it with other components (Figure S5), we further measured the correlation of the MNI-registered weight map for each participant with the average weight map across the 7 other participants (averaged across 8 folds). This analysis showed the highest inter-subject correlation of the weight maps for Component 1 (scenes), followed by Components 2 and 4 (faces and text, respectively), and then Components 3 (food) and 5 (bodies). We next quantified the spatial distribution of the voxel weights for each component using a sparseness measure based on the relationship between L1 and L2 norms.[21] Component 3 was less sparse than the others, except for Component 5. Kurtosis and skewness of the voxel weight distributions of each component show that all components have voxel weights that are positively skewed and kurtotic, relative to a Gaussian, indicating a peakier, heavy-tailed distribution skewed toward higher values. A measure of lateralization of the weight maps showed that the face component was right lateralized and the text component left lateralized, (as expected). The food component trended toward left lateralization, although the lateralization effect was found to be non-significant (one-sample one-tailed t test, $t(8) = 1.76$, $p = 0.06$). Taken together, these analyses indicate that the inter-subject variability, the sparseness, and degree of the lateralization of the food component is within the range of the other components but at the lower end of that range.

Given the small but significant correlation across stimuli of the Component 3 response with various measures of color information, even after food salience was factored out, we next asked how similar the anatomical distribution of this component was to the distribution of color responsiveness across the cortex.[22,23] Specifically, we measured the correlation between the saturation of non-food stimuli with voxel responses in the VVC to those stimuli and then compared the resulting correlation map with the voxel weights for Component 3, separately within each participant. We find (Figure S6) that the saturation-responsiveness map is indeed correlated with the Component 3 weight map in every subject (mean ~ 0.4) and more so than for any other component ($p < 0.01$ for all 4 comparisons using a paired t test). Thus, the anatomical distribution of Component 3 and color responses are correlated with each other across the cortex.

DISCUSSION

We applied data-driven analyses to a very large dataset of fMRI responses to thousands of natural images and found that the dominant neural response profiles in the ventral visual pathway include selective responses to faces, scenes, bodies, text, and food. Although the first four of these selective responses have been reported in many previous studies, what is novel in the current study is their emergence, unbidden, from a hypothesis-neutral analysis of a dataset that was not designed to test for or to reveal them. The fact that these four previously reported selectivities emerged separately within each of the 8 individual participants in this study (each of whom saw mostly non-overlapping images) shows that they reflect not just the idiosyncratic whims of the scientists who chose to test these hypotheses in the past but the actual dominant features of the neural response in the ventral visual pathway. But our most novel result is the discovery of a new neural response that has not been reported previously for the ventral visual pathway and that is highly selective to images of food. Taken together, these results give a more comprehensive and data-driven characterization of the dominant neural response profiles of the ventral visual pathway, describe a new neural selectivity for visual food images, and provide new clues into why we have the neural selectivities we do.

Because our finding of neural selectivity for food was unexpected, we embarked on an extensive series of control experiments to test alternatives to this hypothesis. We found that although the magnitude of response of Component 3 was correlated with the presence of visual features such as color saturation, warm colors, curved shapes, and texture properties, the only factor that remained highly correlated with the food component response when other factors were partialled out was the salience of food in the image (Figure 3A). Second, when we pitted the presence of warm colors against food salience as accounts of the response of this component, we found that food trumped color: cool-colored food produced a higher response in this component than warm-colored non-food. Third, the food selectivity of this component persisted even for computationally matched stimulus pairs (one food, one non-food) that

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

CellPress

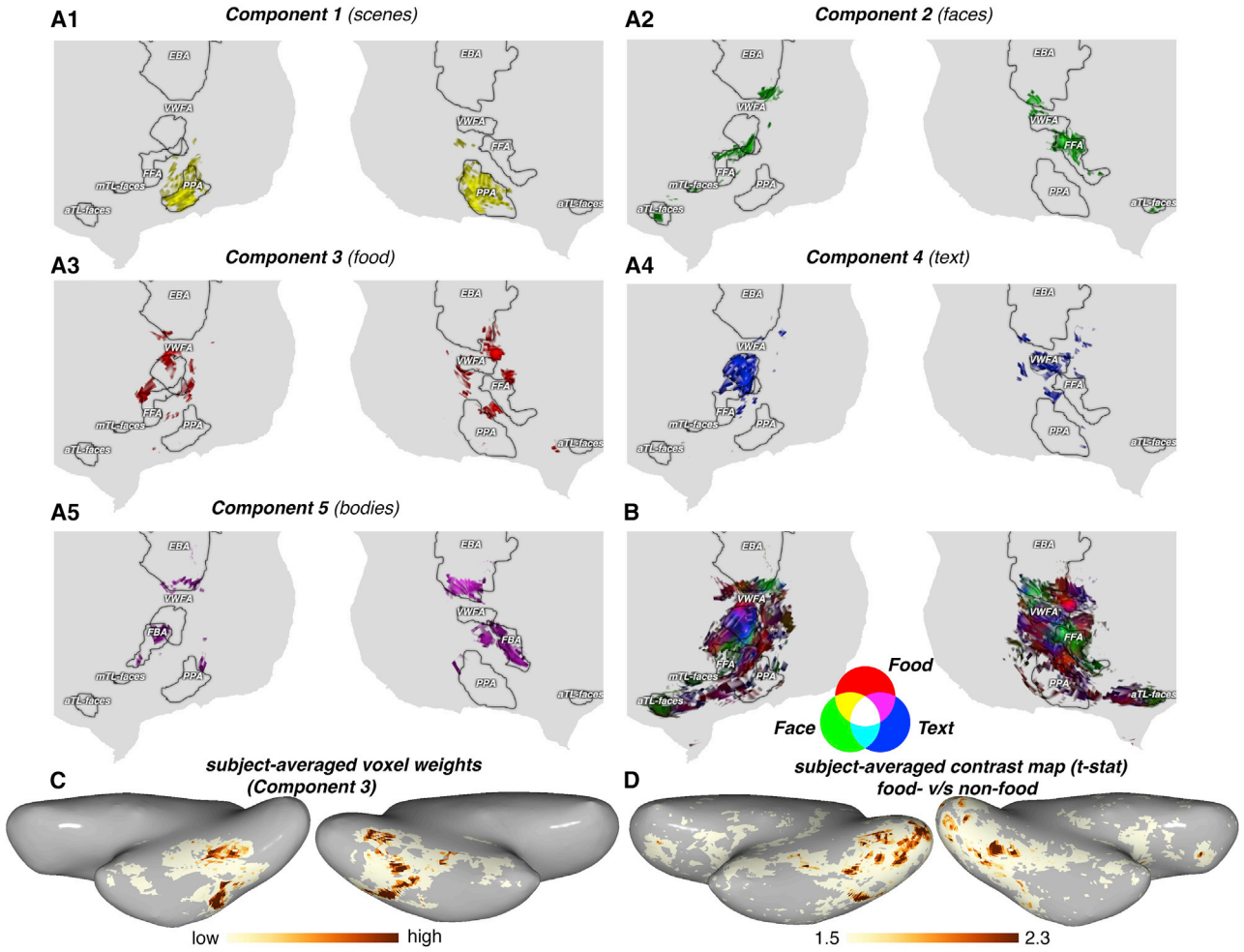

Figure 7. Anatomical locations of the highest weights for each component

(A) Top 5% voxels with the highest weights on each component are visualized on cortical flatmaps for one subject. Established regions of interest, defined from the functional localizer scans by computing the contrast of preferred versus all other stimuli, are shown in outlines (t value > 2.5).

(B) Voxel weight maps of face, food, and text-selective components are visualized together on the RGB colormap to show component overlap for the same subject. Similar maps for the remaining subjects are shown in Video S2.

(C) Component 3 voxel weights for each subject are registered to the MNI space and then averaged across subjects to obtain a subject-averaged voxel weight map.

(D) T-statistics from the food versus non-food contrast experiment on a subset of NSD stimuli (mapped onto MNI space, followed by cross-subject averaging) are shown for comparison with (C).

See also Figures S4 and S5.

elicit similar activation patterns in deep layers of a pre-trained CNN. Fourth, we built a CNN-based model for Component 3,[14] which accurately predicted responses of this component to held-out stimuli, and we used this model to turbocharge our search for counterevidence of the food selectivity of this component. We ran all 1.2 million images from ImageNet through this model and looked at the top 1,000 predicted to produce the highest response. They were all food. Fifth, we constructed pairs of visually similar food and non-food images by hand (e.g., a yellow crescent moon and a banana), and ran these images through our predictive model of Component 3. Again, predicted responses were higher to the food images than to their paired visually similar non-food images. Sixth, an analysis of responses to a subset of nine stimulus categories, including an equal number of stimuli in each category and, importantly, only including food images that are maximally distinct from each other, revealed components selectively responsive to faces and food, but no components responsive to the other 7 categories, showing that the observed selectivity for food is not an artifact of its over-representation in the stimulus set or of the homogeneity of specific kinds of food in the dataset. Taken together, these analyses argue that Component 3 is selectively responsive not to any particular visual features but to food per se. We have therefore labeled this inferred neural population the "ventral food component." Confirming this finding, two preprints based on the same NSD fMRI data, finding voxel-wise selectivity for food and their overlap with color-biased regions, appeared recently.[24,25] Our study using hypothesis-neutral

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

methods further shows that food selectivity is a dominant feature of the ventral visual cortex, that food selectivity is strong when revealed by demixing methods, and that food selectivity overrides low- and mid-level features, including color. One potential caveat of this study is its reliance on complex, naturalistic scenes obtained from crowdsourced data (MS-common objects in context [COCO]).[26] Uncontrolled naturalistic scenes can obscure the relevant image structure that plays a causal role in eliciting neural responses. Follow-up investigations of this component with carefully controlled synthetic stimuli can further clarify the featural drivers of this food selectivity.

While a distinctive response to visual images of food has been described in taste-sensitive regions of the insular cortex,[27] a robust and highly food-selective response in the visual cortex has not been reported before. Why has the food-selective component not been found previously, despite past efforts to look for it?[20,28,29] The likely account is that prior studies primarily analyzed raw voxel responses, which do not reveal strong selectivity for food because the food component is spatially intermingled with other neural populations within voxels. In contrast, our voxel decomposition method demixes these neural responses, revealing the strong selectivity of the food component alone. Indeed, when we obtained the visually matched pairs of food and non-food textures that had produced similar responses in Downing and Kanwisher's study (P. Downing and N. Kanwisher, 1999, Cogn. Neurosci. Soc., poster), leading them to argue against food selectivity, we found that our predictive model for Component 3 produced a higher response to food than non-food. A similar pattern was observed previously with our voxel decomposition analysis of responses in auditory cortex,[8] where we found only weak selectivity for music in raw voxels but strong selectivity for music in the inferred music component, which was later validated by clear music selectivity in the responses of individual intracranial electrodes.[30] Indeed, the precise neural basis of this food-selective population remains a critical open question. The spatially clustered nature of the voxel weights observed for the food component suggests that this inferred food-selective neural population is concentrated in these cortical regions. Finer granularity of neural response measurements, particularly neural recordings of higher spatiotemporal precision that can be achieved with either intracranial measurements in humans or single neuron recordings in macaques (if they have a homologous organization), can further clarify the precise neural basis of this component and shed light on whether each neuron (or electrode site) shows a similar food selectivity or whether a more heterogeneous selectivity is found for different kinds or aspects of food.

A notable property of Component 3 is that even though its selectivity for food cannot be explained by responsiveness to color properties alone, the two are clearly linked. The response profile of Component 3 has a much smaller but still significant correlation with color metrics even after food selectivity has been partialled out, and its anatomical distribution across the cortex is correlated with the anatomical distribution of responsiveness to color information (Figure S6). Why might the apparently same neural population be responsive to both food and color information, even when each is unconfounded from the other? Many have noted the importance of color for the detection, evaluation, and choice of food.[31,32] Neuropsychological studies of patients with cortical color blindness (achromatopsia) have noted particular difficulties in discriminating food. Pallis[33] quotes an achromatopsic patient saying, "I have difficulty in recognizing certain kinds of food on my plate. I can tell peas and bananas by their size and shape. An omelette [sic], however, looks like a piece of meat." Further, behavioral studies have shown that adults, preschool children,[34] and monkeys[35] use color more than shape when generalizing across food categories, but the opposite when generalizing across non-food categories. Indeed, Santos et al.[35] argued that the use of color over shape only in food learning suggests the existence of a domain-specific mechanism for visual food choice. Studies in typical adults also reinforce the deep link between color and food perception, finding, for example, that images of food (but not non-food) are rated as having higher arousal if they contain red and lower arousal if they contain green.[31] The authors of that study speculate that red color was indicative of the caloric and nutritional value of food for our evolutionary ancestors but is much less so today (in prepared foods and where food dyes are used) and hence reveals the evolutionary basis of the connection between color in general, and red in particular, in food preference. Of course, food preferences are famously culture-specific and learned,[36] and an infant's food choice is primarily learned from other people. But color may still play a role in domain-specific learning about food and in bootstrapping the development of a cortical circuit for visual food discrimination. One hypothesis is that the color bias in food choice may arise relatively early in development (though not apparently in infancy[34]), with the cortical locus of the ventral food component accordingly arising in regions already biased for warm colors,[12] but that the particular visual food stimuli that activate this system are most likely learned through individual experience (like orthographies in the VWFA).

What does food selectivity tell us about which categories get their own specialized neural machinery in the brain? Because food has been of fundamental importance to humans both throughout their evolution, and in modern daily life,[36] and because food choice often starts with vision, a specialization for food in the visual cortex is consistent with both evolutionary and experiential origins of cortical specializations. On the other hand, food seems more visually heterogeneous than other categories with selective responses in the ventral pathway, an impression confirmed by visual similarity measures based on feature responses in pre-trained AlexNet (Figure S6). Nonetheless, food is linked to some visual features, notably color, and indeed we find that Component 3 does show a small but significant color preference, even after food salience is partialled out. This finding is reminiscent of other feature biases in category-selective cortex (e.g., curvature biases in face-selective cortex[37]), and invites the same chicken-and-egg question: do category selectivities colonize cortical regions with pre-existing relevant feature biases,[37] or are these visual feature preferences simply by-products of category selectivity? Finally, the finding of food selectivity resolves a previous conflict with the hypothesis that category selectivity in visual cortex is determined by the computational requirements of the task.[38] We had proposed this hypothesis in a recent study,[38] based on our finding that convolutional neural networks trained on both face discrimination and object classification spontaneously segregated themselves into separate systems for face and object recognition. But that study also found

spontaneous segregation for food in a network trained on both food and object classification, a finding that seemed then not to fit the brain but that now does. Thus, the novel selectivity for food reinforces the computational hypothesis that task constraints play a role in determining which categories are processed with their own specialized neural machinery.

What computational advantages might a selective response to food confer? Any form of selectivity in a neural population inherently implies a sparse code, as it suggests that the neural population responds strongly to only a specific subset of all possible stimuli. Such sparse neural codes have long been argued to make information explicit and easier to read out[39–41] and to support faster learning.[42–44] Reducing metabolic costs would also favor sparse codes for stimuli that are most frequently encountered in our environment. Of course, we cannot have specialized neural codes for all possible classes of stimuli, so it would be sensible to allocate such specialized systems to a relatively small number of the most important object classes—such as food.

A final note is that our analysis did not find evidence for selective neural responses to several visual features and categories for which ventral visual pathway specializations have been proposed in the past, including animals[45] (which are well represented in the stimulus set) and stubby-shaped and spikey-shaped objects.[46] We also did not see evidence for previously proposed selectivities for small inanimate objects[45,47] or tools,[48,49] although these selectivities may be located more on the lateral than ventral surface of the brain, outside the search window used here. Of course, there are many reasons why selectivities that exist in the brain might not be detected using fMRI, but the failure of previous findings from fMRI to emerge from the current analysis raises questions about whether those selectivities might be better accounted for by the components found here.[50]

In sum, our hypothesis-neutral investigation on the ventral visual pathway reveals neural populations selective for faces, scenes, bodies, text, and food. The fact that these selectivities emerge from a hypothesis-neutral analysis, across multiple largely non-overlapping sets of images, indicates that they reflect not just the capricious interests of researchers in the past but rather constitute dominant features of the functional organization of the ventral visual pathway. Further, the novel selectivity for food reported here raises fascinating questions about its developmental origins, connectivity, and behavioral consequences. Another important open question is whether this neural population represents the mere presence of food, or its familiarity, appeal, or caloric or nutritive content. This new finding further shows that selective neural responses in the ventral visual pathway arise not only for perceptually homogeneous categories that may reflect confluences of overlapping visual feature maps[51] but also categories that are visually quite heterogeneous, especially if exemplars of that category require specialized computations for their discrimination[38] and engage our frequent and abiding interest.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Natural scenes dataset
  - BOLD5000
- METHOD DETAILS
  - Data selection criterion and preprocessing
  - Independent replication with held-out subjects data
  - Online behavioral experiment
  - Affective ratings from NSD-meadows behavioral dataset
  - Sampling matched food and non-food stimuli
  - Sampling diverse subsets of food and non-food stimuli
  - Curated subset of Natural Scenes Dataset with balanced categories
  - Sampling food and non-food stimuli for the food contrast experiment
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Component modeling
  - Component selectivity analysis
  - Control analysis on the novel component
  - Analysis on curated Natural Scenes Dataset
  - Analysis on the independent BOLD5000 dataset
  - Quantification and statistical analysis on all component voxel weights
  - Food contrast experiment
  - Encoding model of the inferred components

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2022.08.009.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant DP1HD091947 to N.K., K99/R00 Pathway to Independence Award from the National Eye Institute of the NIH (grant 1K99EY032603) to N.A.R.M., and the Center for Brains, Minds, and Machines (CBMM) funded by NSF STC award CCF-1231216. Collection of the NSD dataset was supported by NSF IIS-1822683 and NSF IIS-1822929. This work was presented on May 15, 2022, at the Vision Sciences Society (abstract submitted December 2021). We thank Elizabeth Meiczkowski for help obtaining behavioral data and members of the Kanwisher lab for helpful suggestions on the manuscript.

AUTHOR CONTRIBUTIONS

N.K. and M.K. conceived the study. M.K. performed the main NMF analyses in the paper and N.A.R.M. built the CNN model and performed analysis on its predictions. N.K., M.K., and N.A.R.M. wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 23, 2022
Revised: August 3, 2022
Accepted: August 5, 2022
Published: August 25, 2022

REFERENCES

1. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nat. Neurosci. *25*, 116–126.
2. Kanwisher, N., and Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. Philos. Trans. R. Soc. Lond. B Biol. Sci. *361*, 2109–2128.
3. Epstein, R., and Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature *392*, 598–601.
4. Polk, T.A., and Farah, M.J. (1998). The neural development and organization of letter recognition: evidence from functional neuroimaging, computational modeling, and behavioral studies. Proc. Natl. Acad. Sci. USA *95*, 847–852.
5. Schalk, G., Kapeller, C., Guger, C., Ogawa, H., Hiroshima, S., Lafer-Sousa, R., Saygin, Z.M., Kamada, K., and Kanwisher, N. (2017). Facephenes and rainbows: causal evidence for functional and anatomical specificity of face and color processing in the human brain. Proc. Natl. Acad. Sci. USA *114*, 12285–12290.
6. Downing, P.E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. Science *293*, 2470–2473.
7. Pitcher, D., Charles, L., Devlin, J.T., Walsh, V., and Duchaine, B. (2009). Triple dissociation of faces, bodies, and objects in extrastriate cortex. Curr. Biol. *19*, 319–324.
8. Norman-Haignere, S., Kanwisher, N.G., and McDermott, J.H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron *88*, 1281–1296.
9. Schmidt, M.N., Winther, O., and Hansen, L.K. (2009). Bayesian non-negative matrix factorization. In Lecture Notes in Computer Science. International Conference on Independent Component Analysis and Signal Separation (Springer), pp. 540–547.
10. Khosla, M., Ratan Murty, N.A., and Kanwisher, N. (2022). Data-driven component modeling reveals the functional organization of high-level visual cortex. Sci Rep. *7*, 3596.
11. Grill-Spector, K., and Weiner, K.S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. Nat. Rev. Neurosci. *15*, 536–548.
12. Rosenthal, I., Ratnasingam, S., Haile, T., Eastman, S., Fuller-Deets, J., and Conway, B.R. (2018). Color statistics of objects, and color tuning of object cortex in macaque monkey. J. Vis. *18*, 1.
13. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. *25*.
14. Ratan Murty, N.A., Bashivan, P., Abate, A., DiCarlo, J.J., and Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. Nat. Commun. *12*, 5540.
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. Preprint at arXiv. https://doi.org/10.48550/arXiv.2103.00020.
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009). Imagenet: a large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.
17. Freeman, M. (2019). Time use of millennials and nonmillennials. Mon. Labor Rev. 1–13.
18. Chang, N., Pyles, J.A., Marcus, A., Gupta, A., Tarr, M.J., and Aminoff, E.M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. Sci. Data *6*, 49.
19. Prince, J.S., Charest, I., Kurzawski, J.W., Pyles, J.A., Tarr, M.J., and Kay, K.N. (2022). GLMsingle: a toolbox for improving single-trial fMRI response estimates. Preprint at bioRxiv. https://doi.org/10.1101/2022.01.31.478431.
20. Downing, P.E., Chan, A.W., Peelen, M.V., Dodds, C.M., and Kanwisher, N. (2006). Domain specificity in visual cortex. Cereb. Cortex *16*, 1453–1461.
21. Hoyer, P.O. (2004). Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. *5*, 1457–1469.
22. Lafer-Sousa, R., Hermann, K.L., and Conway, B.R. (2015). Striking individual differences in color perception uncovered by 'the dress' photograph. Curr. Biol. *25*, R545–R546.
23. Conway, B.R. (2018). The organization and operation of inferior temporal cortex. Annu. Rev. Vis. Sci. *4*, 381–402.
24. Jain, N., Wang, A., Henderson, M.M., Lin, R., Prince, J.S., Tarr, M.J., and Wehbe, L. (2022). Food for thought: selectivity for food in human ventral visual cortex. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.22.492983.
25. Pennock, I.M.L., Racey, C., Allen, E.J., Wu, Y., Naselaris, T., Kay, K., Franklin, A., and Bosten, J. (2022). Color-biased regions in the ventral visual pathway are food-selective. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.25.493425.
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C.L. (2014). Microsoft COCO: common objects in context. In Lecture Notes in Computer Science. European Conference on Computer Vision (Springer), pp. 740–755.
27. Avery, J.A., Liu, A.G., Ingeholm, J.E., Gotts, S.J., and Martin, A. (2021). Viewing images of foods evokes taste quality-specific activity in gustatory insular cortex. Proc. Natl. Acad. Sci. USA *118*. e2010932118.
28. Pohl, T.M., Tempelmann, C., and Noesselt, T. (2017). How task demands shape brain responses to visual food cues. Hum. Brain Mapp. *38*, 2897–2912.
29. van der Laan, L.N., De Ridder, D.T., Viergever, M.A., and Smeets, P.A. (2011). The first taste is always with the eyes: a meta-analysis on the neural correlates of processing visual food cues. Neuroimage *55*, 296–303.
30. Norman-Haignere, S.V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J.H., Schalk, G., and Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. Curr. Biol. *32*, 1470–1484.e12.
31. Foroni, F., Pergola, G., and Rumiati, R.I. (2016). Food color is in the eye of the beholder: the role of human trichromatic vision in food evaluation. Sci. Rep. *6*, 37034.
32. Spence, C. (2015). On the psychological impact of food colour. Flavour *4*, 1–16.
33. Pallis, C.A. (1955). Impaired identification of faces and places with agnosia for colours: report of a case due to cerebral embolism. J. Neurol. Neurosurg. Psychiatry *18*, 218–224.
34. Shutts, K., Condry, K.F., Santos, L.R., and Spelke, E.S. (2009). Core knowledge and its limits: the domain of food. Cognition *112*, 120–140.
35. Santos, L.R., Hauser, M.D., and Spelke, E.S. (2001). Recognition and categorization of biologically significant objects by rhesus monkeys (Macaca mulatta): the domain of food. Cognition *82*, 127–155.
36. Rozin, P., and Todd, P.M. (2016). The evolutionary psychology of food intake and choice. In The Handbook of Evolutionary Psychology: Foundations (John Wiley & Sons), pp. 183–205.
37. Arcaro, M.J., and Livingstone, M.S. (2021). On the relationship between maps and domains in inferotemporal cortex. Nat. Rev. Neurosci. *22*, 573–583.
38. 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.
39. Barlow, H.B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? Perception *1*, 371–394.
40. Földiak, P., and Young, M.P. (1995). Sparse coding in the primate cortex. In Handbook of Brain Theory and Neural Networks (MIT Press), pp. 1–1064.
41. Olshausen, B.A., and Field, D.J. (2004). Sparse coding of sensory inputs. Curr. Opin. Neurobiol. *14*, 481–487.

42. Fiete, I.R., Hahnloser, R.H., Fee, M.S., and Seung, H.S. (2004). Temporal sparseness of the premotor drive is important for rapid learning in a neural network model of birdsong. J. Neurophysiol. *92*, 2274–2282.
43. Márton, C.D., Zhou, S., and Rajan, K. (2022). Linking task structure and neural network dynamics. Nat. Neurosci. *25*, 679–681.
44. Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F., and Ostojic, S. (2022). The role of population structure in computations through neural dynamics. Nat. Neurosci. *25*, 783–794.
45. Konkle, T., and Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. J. Neurosci. *33*, 10235–10242.
46. Bao, P., She, L., McGill, M., and Tsao, D.Y. (2020). A map of object space in primate inferotemporal cortex. Nature *583*, 103–108.
47. Magri, C., Konkle, T., and Caramazza, A. (2021). The contribution of object size, manipulability, and stability on neural responses to inanimate objects. NeuroImage *237*, 118098.
48. Chao, L.L., Haxby, J.V., and Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nat. Neurosci. *2*, 913–919.
49. He, C., Hung, S.C., and Cheung, O.S. (2020). Roles of category, shape, and spatial frequency in shaping animal and tool selectivity in the occipitotemporal cortex. J. Neurosci. *40*, 5644–5657.
50. Ritchie, J.B., Zeman, A.A., Bosmans, J., Sun, S., Verhaegen, K., and Op de Beeck, H.P.O. (2021). Untangling the animacy organization of occipitotemporal cortex. J. Neurosci. *41*, 7103–7119.
51. Op de Beeck, H.P., Haushofer, J., and Kanwisher, N.G. (2008). Interpreting fMRI data: maps, modules and dimensions. Nat. Rev. Neurosci. *9*, 123–135.
52. Pnevmatikakis, E.A., Soudry, D., Gao, Y., Machado, T.A., Merel, J., Pfau, D., Reardon, T., Mu, Y., Lacefield, C., Yang, W., et al. (2016). Simultaneous denoising, deconvolution, and demixing of calcium imaging data. Neuron *89*, 285–299.
53. Kotliar, D., Veres, A., Nagy, M.A., Tabrizi, S., Hodis, E., Melton, D.A., and Sabeti, P.C. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-seq. eLife *8*, e43803.
54. Hasler, D., and Suesstrunk, S.E. (2003). Measuring colorfulness in natural images. In Proceedings of SPIE - The International Society for Optical Engineering, *5007*, pp. 87–95.
55. Li, S.P.D., and Bonner, M. (2020). Curvature as an organizing principle of mid-level visual representation: a semantic-preference mapping approach. In 2nd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM), NeurIPS 2020. https://openreview.net/pdf?id=CUi1G2UWsAm.
56. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 586–595.
57. Dosovitskiy, A., and Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. Adv. Neural Inf. Process. Syst. *29*.
58. Lawson, C.L., and Hanson, R.J. (1995). Solving Least Squares Problems (SIAM).
59. Lafer-Sousa, R., Conway, B.R., and Kanwisher, N.G. (2016). Color-biased regions of the ventral visual pathway lie between face- and place-selective regions in humans, as in macaques. J. Neurosci. *36*, 1682–1697.

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
fMRI dataset of responses in 8 subjects to 9,000-10,000 stimuli, Allen et al.[1]	Natural Scenes Dataset[1]	http://naturalscenesdataset.org/
fMRI dataset of responses in 4 subjects to ~5,000 stimuli	BOLD5000 v2[18,19]	https://bold5000-dataset.github.io/website/download.html
ImageNet dataset	ImageNet: A large-scale hierarchical image database[16]	https://www.image-net.org/challenges/LSVRC/
MSRA Image Set: Dataset used for extracting color statistics (object color probability)	Microsoft Research Asia, MSRA[12]	https://neicommons.nei.nih.gov/#/objectcolorstatistics
Software and algorithms		
Pre-trained AlexNet	AlexNet Model[13]	https://pytorch.org/hub/pytorch_vision_alexnet/
OpenAI Clip Model	OpenAI Github release[15]	https://github.com/openai/CLIP
NIMFA package	Python library	https://github.com/mims-harvard/nimfa
Pycortex	Jack Gallant lab at UC Berkeley	https://github.com/gallantlab/pycortex; https://doi.org/10.3389/fninf.2015.00023
Other		
Downing Stimuli	Paul Downing, Kanwisher lab (P. Downing and N. Kanwisher, 1999, Cogn. Neurosci. Soc., poster)	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for code and resources should be directed to and will be fulfilled by the lead contact, Meenakshi Khosla (mkhosla@mit.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The paper analyzed existing, publicly available data. The original dataset on which the component decomposition was performed is listed in the key resources table and is available as of the data of publication at: http://naturalscenesdataset.org/.
- All original code has been deposited at this github repository: https://github.com/mk2299/ComponentModeling.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Natural scenes dataset

A detailed description of the Natural Scenes Dataset (NSD; http://naturalscenesdataset.org) is provided in the original study that collected the dataset (Allen et al.[1]). Briefly, the NSD contains measurements of fMRI responses from 8 participants (S1-8), who each viewed 9,000–10,000 distinct color natural scenes (22,000–30,000 trials) over the course of 30–40 scan sessions. Subject demographics are as follows: S1: male, age 30; S2: female, age 28; S3: female, age 29; S4: female, age 27; S5: female, age 32; S6: male, age 23; S7: female, age 24; S8: female, age 19. Scanning was conducted at 7T using whole-brain gradient-echo EPI at 1.8-mm resolution and 1.6-s repetition time. Images were taken from the Microsoft Common Objects in Context (COCO) database,[26] square cropped, and presented at a size of 8.4° x 8.4°. A special set of 1,000 images were shared across half the subjects with full repetitions (participant NSD IDs: 1,2,5,7) and a subset of these (515 images) were shared across all 8 participants. The remaining images were mutually exclusive across subjects. Images were presented for 3 s with 1-s gaps in between images. Subjects fixated centrally and performed a long-term continuous recognition task on the images. The fMRI data were pre-processed by performing one temporal

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

interpolation (to correct for slice time differences) and one spatial interpolation (to correct for head motion). A general linear model was then used to estimate single-trial beta weights. Cortical surface reconstructions were generated using FreeSurfer, and both volume- and surface-based versions of the beta weights were created.

BOLD5000

We also analyzed another publicly available large-scale dataset, namely, the BOLD5000 dataset.[18,19] This dataset comprises BOLD responses from four participants (CSI1-4), while they each viewed nearly 5,000 natural images, though most images had only single repetitions. Subject demographics are as follows: CSI1: male, age 27; CSI2: female, age 26; CSI3: female, age 24; CSI4: female, age 25.

METHOD DETAILS

Data selection criterion and preprocessing

Natural scenes dataset

In this paper, we used the 1.8-mm volume 'nativesurface' preparation of the NSD data and version 3 of the NSD single-trial betas (betas_fithrf_GLMdenoise_RR). We analyzed responses only to images that were seen three times in the participant in question. This leads to 10,000 images and corresponding brain measurements in Phase I subjects (NSD IDs: 1, 2, 5 and 7) and 5,445 images and corresponding brain data in Phase II subjects (NSD IDs: 3, 4, 6, 8).

We averaged single-trial betas across the 3 repetitions after z-scoring every voxel separately within each scan session to create our voxel responses. This within-scan normalization was performed to account for differences in mean percent signal change (PSC) across scan sessions, which may arise due to incidental variability in global BOLD signals.

We extracted ventral visual stream voxels by using the streams atlas provided in the native space of each subject with NSD. Briefly, this ROI collection reflects large-scale divisions of the visual cortex into primary visual cortex and intermediate and high-level ventral, lateral and dorsal visual areas. These were manually drawn for each subject by NSD curators and were based on the voxel-level reliability metrics. For this study, we extracted the ROI mask corresponding to the 'higher-level ventral stream' label. This ROI was drawn to follow the anterior lingual sulcus (ALS), including the anterior lingual gyrus (ALG) on its inferior border and to follow the inferior lip of the inferior temporal sulcus (ITS) on its superior border. The anterior border was drawn based on the midpoint of the occipital temporal sulcus (OTS). As shown in Figure 1A, it is very broad (~7,000-9,000 voxels per subject).

To make the data matrix suitable for NMF so that it contains all positive entries, we perform a baseline shift of voxel responses by subtracting the minimum z-scored response of each voxel (across all stimuli) from its responses to all stimuli.

BOLD5000

In this study, we analyzed the release 2.0 of the BOLD5000 data that integrates custom hemodynamic response function estimation and GLM denoising (data version descriptor: TYPED-FITHRF-GLMDENOISE-RR). Here, we restrict our focus to responses for stimuli that had food/no-food annotations in the BOLD5000 dataset, namely, the MS-COCO images. This includes BOLD activations for the shared set of 1,000 images viewed by all NSD participants and a novel set of 412 images that were unique to BOLD5000 and were viewed at least once by each of the BOLD5000 participants.

Independent replication with held-out subjects data

The hypothesis that the ventral visual pathway contains a neural population that responds selectively to food was formulated based on analyses of the data from four participants in Phase 1 (NSD participants 1,2,5,7), before hypotheses and analysis methods were registered on OSF, and then tested on the held-out data in Phase 2 (NSD participants 3,4,6,8).

Only the top few images for the 5 most inter-subject consistent components were visually inspected to ascribe semantic categories to components. The same top 5 components (with top images respectively selective for faces, scenes, food, text, and bodies) were obtained in the Phase 2, confirming reproducibility of our findings and allowing us to combine responses across the different phases. This apparent category selectivity of each of the top 5 components was subsequently rigorously assessed using quantitative measures based on salience ratings, as described below.

Online behavioral experiment

In an online experiment, we collected subjective salience ratings from Amazon Mechanical Turk for the 1,000 images viewed by all the Phase I participants. Among these images, 515 images were also viewed by all Phase II participants with three repetitions. In this experiment, participants were asked to rate the salience of each of the 5 categories that seemed to be intuitively represented in the top images for each component, namely, *scenes, faces, bodies, text*, and *food*. Specifically, participants were given the following task: 'Rate how prominent [category] is within each image' and were instructed to provide a rating on a scale of 0 to 9. Each participant completed salience ratings for one category over a series of 220 images and we obtained 5 ratings per category (from 5 participants) and averaged the ratings across participants to get an overall measure of the salience of each category for each image. Salience ratings for scenes were odd, presumably because most natural images have some kind of scene context, and participants were unsure what we meant. Therefore, for the scene category only, we instead asked two experts in scene-selective cortex who were uninvolved in the study to rate their prediction for how strongly the image would drive the scene-selective cortex.

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

Affective ratings from NSD-meadows behavioral dataset

Valence and arousal ratings were obtained from the NSD Meadows behavioral dataset which was released along with the NSD dataset (Further details provided in Allen et al.[1]). These ratings were released for a total of 100 images from the shared image set viewed by all NSD participants.

Sampling matched food and non-food stimuli

We identified pairs of images (one food and one nonfood) from the 5,445-10,000 image set viewed by each participant that produce similar activations in the last convolutional layer ('conv5') of an AlexNet[13] pre-trained on ImageNet[16] categorization. This matching analysis was performed on the entire image set of each participant at the subject-level (not just shared images), since having a larger stimulus set increases the chance of finding a stronger control pair. This computational matching procedure yielded 20-40 image pairs per participant. Mean food-non food similarity score, computed as the correlation between the features in the conv5 layer for the food image and the corresponding non-food image, across all pairs was thus very high (mean = 0.55, s.d. = 0.09); importantly, all pairs had a similarity score > 0.33. This was further substantially greater than the pairwise similarity among all the food images (mean = 0.16, s.d. = 0.17) and non-food images separately (mean = 0.11, s.d. = 0.18). Visual inspection further confirmed that these matched food and non-food images contain similar visual features like similar colors and textures (example pairs shown in Figure 3C).

Sampling diverse subsets of food and non-food stimuli

We selected diverse subsets of food and non-food stimuli that maximally span the representational space of different layers of a pre-trained CNN, such that the images within each subset are substantially more dissimilar (in terms of the average pairwise distance computed in the representational space of the corresponding layer) than what would be expected if the images were drawn at random from the respective set. The images were selected greedily to maximize the distance of each image with its closest neighbor.

The procedure is outlined as follows: for each layer l,

(i) we first randomly sample a food image,
(ii) we then select the next image from the set of all food images viewed by each participant (n=10,000 in Phase 1 and 5,445 in Phase 2) as the image which has the largest correlation distance (1-r) to its closest neighbor among the already selected food images, where the distance is computed between the image features extracted at layer l and
(iii) we repeat (ii) until we get the desired number of images (n=50).

The same procedure is repeated for the set of non-food images as well to get n=50 non-food images. These selected images are so diverse that a linear classifier trained to discriminate between these food and nonfood images using the features of the corresponding layer l performs at chance (never exceeding 53% across all layers), presumably because there is no remaining simpler visual characteristic shared by stimuli within the two subsets that a classifier can latch onto.

Curated subset of Natural Scenes Dataset with balanced categories

To test whether a high proportion of exemplars of any category in the dataset might be sufficient for a component to emerge that responds selectively to that category, we sampled a subset of images with a fixed number of examples from each of 9 categories. These categories were selected because there were enough images in each subject belonging to those categories and include the following: face, food, clock, airplane, elephant, giraffe, horse, truck, motorcycle. Importantly, we chose an equal number of stimuli for each of these categories within each subject in this subset (although this number varied slightly across subjects because each subject saw different images, n=197, 207, 182, 191 per category for Phase 1 participants 1, 2, 5 and 7, respectively). The food images in this subset were drawn so that they are maximally heterogenous (dissimilar amongst themselves) following the procedure for sampling diverse subsets described above.

Sampling food and non-food stimuli for the food contrast experiment

We selected a subset of food and non-food images from the shared set of 515 images viewed by all 8 NSD participants. We selected all images that had a food salience rating > 4 for the 'food' category and sampled an equal number of non-food images (food salience rating = 0), yielding 49 food and non-food images. These images and their corresponding brain responses were used to emulate a food contrast experiment.

QUANTIFICATION AND STATISTICAL ANALYSIS

Component modeling

A Bayesian Matrix Factorization approach for the analysis of large-scale fMRI recordings

We model the data matrix (voxels x images) as the product of two lower rank matrices. The first matrix (called the response profile matrix henceforth) encodes the response profiles of each component ('neural populations') to all images and the second matrix (called the component by voxel weight matrix) specifies the relative contribution of all voxels to each component. We chose NMF for our matrix factorization algorithm for several reasons. First, PCA/ICA based approaches do not yield ''signed'' components, i.e., negative and positive weights are treated equivalently. Initial pilot analyses of our data using the PCA/ICA approach of

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

Norman-Haignere et al.[8] revealed single components with both positive and negative responses and voxel weights which couldn't be oriented (or flipped) such that the response and voxel weights predominantly have the same sign. Negative response magnitudes are generally not consistent with neural responses in the ventral visual pathway, which usually increase after stimulus presentation, and negative voxel weights violate our modeling assumptions about the voxel weight matrix representing the relative anatomical proportions of each component in every voxel. Further, ICA requires the statistical independence of unmixed components and can fail in practice when no linear demixing matrix is found, as can happen when there is significant spatial overlap between distinct neural populations and independence is not achievable. NMF is better equipped to handle spatially overlapping signals in such cases, and has been more effective in other neuroscience domains that rely on demixing of spatially overlapping components.[52] Importantly, NMF makes only minimal and biologically meaningful assumptions about the components by enforcing the basis functions to be nonnegative. These considerations led us to favor an NMF-based approach over other decomposition techniques. Since choosing the number of components is an important problem in NMF, we adopted a Bayesian NMF approach[9] since it affords a principled way of selecting the number of components based on likelihood (e.g., Bayesian Information criterion). For a comparison of the food selectivity of the NMF-derived food component with the most food-selective PCA component and cluster (given by k-means clustering), see Figure S7.

Mathematically, the Bayesian NMF algorithm models the data matrix **D** as,

$$\boldsymbol{D} = \boldsymbol{RV} + \boldsymbol{E},$$

where **D** is the images x voxels data matrix for every participant, **R** is the image x components (N x C) response profile matrix, **V** is the components x voxels (C x V) voxel weight matrix and **E** is an images x voxels (N x V) residual matrix. In the Bayesian approach to NMF, all parameters for (**R, V, E**) are stated in terms of their prior densities. For efficient inference, following Schmidt et al.,[9] we choose a zero mean normal residual matrix **E** with variance σ^2, and a normal data likelihood,

$$p(\boldsymbol{R}, \boldsymbol{V}, \boldsymbol{E}) \sim \Pi_{n=1,..,N;v=1,..,V} \mathcal{N}\left(\boldsymbol{D}_{n,v}; (\boldsymbol{RV})_{n,v}, \sigma^2\right).$$

Further we assume that R and V are independently exponentially distributed with scales $\rho_{n,c}$ and $\gamma_{c,v}$,

$$p(\boldsymbol{R}) \sim \Pi_{n=1,..,N;c=1,..,C}\, \rho_{n,c} \exp\left(-\rho_{n,c} R_{n,c}\right) 1(R_{n,c} > 0)$$

$$\text{and } p(\boldsymbol{V}) \sim \Pi_{c=1,..,C;v=1,..,V}\, \gamma_{c,v} \exp\left(-\gamma_{c,v} V_{c,v}\right) 1(V_{c,v} > 0)$$

The conditional probabilities of **R** and **V** thus have a rectified Gaussian distribution. Following Schmidt et al.,[9] the prior for the variance in **E** is assumed to have an inverse gamma distribution, resulting in an inverse-gamma conditional probability. Parameters for (**R, V, E**) are optimized by sequentially drawing samples from these conditional densities using the Bayesian Markov Chain Monte Carlo (MCMC) sampling method derived in Schmidt et al.[9]

Extracting robust components in individual subjects with a consensus approach

Like standard NMF, Bayesian NMF is also a stochastic algorithm sensitive to initialization and accurate initialization of the estimates is critical. To get robust components, we run this algorithm N= 50 times on the data matrix for each subject to get C=20 components per run. We then perform a consensus NMF procedure inspired by Kotliar et al.[53] to aggregate results from different runs of the NMF algorithm into a single stable matrix factorization result. In this procedure, the estimated response profile matrices from each run are concatenated across the component dimension to create an (images x NC) matrix where each column is a component from a single run of the algorithm. We follow the same procedure as described in Kotliar et al.[53] to get the consensus response profile matrix (images x C) from this aggregated data matrix. This consensus algorithm first isolates and removes unreliable components by running an outlier detection procedure, enabling us to filter out components that are not replicable across runs. Next, the remaining components over all runs combined are clustered (with C clusters) and the medians of these clusters are returned as the consensus (stable) response profiles of the C components.

The final voxel weight matrix for each subject is then obtained by finding component indices in individual NMF runs that have the highest correlation with each of the C consensus NMF component response profiles. The respective voxel weights for each index are normalized (to sum up to 1) and then averaged across runs. This gives us the consensus voxel weights for each component.

Extracting components with high inter-subject consistency

The previous analysis yielded 20 components in each individual subject. Since we are interested in discovering the functional organization structure shared across individuals, we next analyzed the one-to-one correspondence between these components across subjects. To determine which of the resulting components for each participant are shared across participants, we use the 1,000 images (or 515 images in the case of Phase II participants) that were viewed by all participants. Specifically, we rank-ordered components based on the highest average inter-subject correlation in their response to the shared images. Since there are 4 subjects in each phase of our analysis, we get 6 unique pairwise correlation values for every possible combination of ordered component indices across the 4 subjects (20 x 20 x 20 x 20). The inter-subject correlation measure, called 'inter-subject consistency', is computed as the average of these 6 values. We first pick the component indices (i, j, k, l) that yield the highest inter-subject consistency. We then repeat the same procedure on the (19 x 19 x 19 x 19) matrix after removing the indices (i, j, k, l) and repeat this procedure until the inter-subject correlation drops significantly. As shown in Figure 1, this

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

value drops sharply after a handful of components, and we restrict our analysis to the top 5 components which all demonstrate an average inter-subject correlation value of 0.5 or greater. See also Figure S1.

Component selectivity analysis

We performed a correlation analysis (Pearson's r) to quantify the extent of agreement between the responses of each component and the salience ratings of their preferred category (as visualized in the top images) over all 515 images that were viewed by all participants. Component responses were first averaged across all 8 subjects before correlation computation (Figure 2). We further also computed these correlations at the single-subject level using all stimuli for which the salience ratings were available (1,000 for Phase I participants and 515 for Phase II participants), as shown in Figure S2.

Control analysis on the novel component

Response correlations with lower-level image-level properties

We computed the following image-level properties to assess their respective impact on driving Component 3's responses (our alternative accounts for food selectivity). These properties include:

Color metrics:

(a) Saturation: Mean saturation of every image is computed after transforming the image from RGB to HSV space
(b) Brightness: Brightness is computed as the mean value across the 'V' channel after transforming the image from RGB to HSV color space.
(c) Colorfulness: This metric is included to capture the *perception* of colorfulness. We compute the colorfulness metric for every image based on the opponent color space representation discussed in Hasler and Suesstrunk.[54]
(d) Hues (Redness): The histogram of the hue channel is computed after binning the hue values across all spatial locations in the image into 8 equally spaced radial bins. The top hue (among the 8 bins) that had the highest correlations with the food-selective component's response roughly corresponded to red hues. We thus included the hue values in this bin in the subsequent partial correlation analysis while assessing the unique contribution of each metric in explaining this component's response.
(e) A color representational axis defined in Rosenthal et al.,[12] called 'Object-color probability' is computed as the probability of a given hue being a natural object in an image. Using the natural image database of over 20,000 images annotated with object segmentation masks (data curated by Microsoft and further annotated and analyzed in Rosenthal et al.[12]), we computed the object probability for each color using the procedure described in Rosenthal et al.[12] as follows: (i) Each image is first encoded in the cylindrical representation of the Lu'v' chromaticity space, namely the Hue-Chroma-Luminance color space (ii) Number of natural object and background pixels that fall within each color bin (from 240 colors bins at 24 equally spaced hue and 10 equally space chromas values) are then computed separately using the segmentation masks of natural objects. (ii) The object probability of each color is then derived as the number of pixels having that color in natural objects divided by the number of pixels having the same color in either natural objects or background. Once the probabilities are estimated, we compute the mean object color probability for each NSD image as the average of the probability over all color bins weighted by the number of pixels in the image that fall within each color bin.

Texture:

We use entropy as a loose local statistical measure for texture. Entropy (**E**) is computed as the Shannon's entropy of the grayscaled version of every image.

$$\mathbf{E} = -\sum p_k \log p_k$$

where p_k is the probability of pixels to have a grayscale intensity value of k.

Curvature index:

We used an image-computable curvature index to estimate the average curvature of contours in every image (as implemented by Li and Bonner[55]). This model convolves the grayscale version of each stimulus with a curvature filter bank with 176 different filters (16 orientations and 11 levels). Each filter in this bank functions as a curved contour detector with a specific orientation and curvature level. The grayscale image is also fed to an edge detection algorithm to find the edge pixels in each image. The overall curvature index is finally estimated by taking the average curvature over all the edge pixels in the image.

To quantify the relationship between these image-computable properties and Component 3 responses, we performed two analyses:

Correlation analysis. We measured the relationship between each of the above variables and the responses of Component 3 to the shared image set using Pearson's correlation coefficient.

Partial correlation analysis. We also performed a partial correlation analysis to assess the unique variance explained by each of the above image-level metrics in the responses of Component 3. We computed the correlation of the residuals resulting from a linear regression of all the above variables individually and food-salience ratings on the responses of Component 3 (the food-selective component). For food, we partialled out the effect of all the above confounders while computing the partial correlation.

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

Analysis on computationally matched food and non-food pairs

We identified pairs of images (one food and one nonfood) from the 5,445-10,000 image set viewed by each participant that produce similar activations in the last convolutional layer ('*conv5*') of an AlexNet[13] pre-trained on ImageNet[16] categorization, as described above. We performed a paired t test to compare the responses of Component 3 to these food-nonfood pairs, separately for each participant. Mean responses of Component 3 to these matched stimuli (averaged across the food and non-food categories per subject) are further shown in Figure 3C.

Analysis on warm-colored non-food and cool-colored food stimuli

Since the partial correlation analysis revealed a low, yet significant correlation of Component 3 responses with the object color probability measure (which reflects the warm-cool color continuum), we performed a subsequent analysis by directly pitting food preference against warm-color preferences. We sampled 50 food images from the lowest end of the object-color probability distribution over all 5,445-10,000 images (bottom 15 percentile) per participant and sampled non-food stimuli from the highest end of this distribution (top 15 percentile). This resulted in cool-colored food stimuli and warm-colored non-food stimuli. We then compared the responses of Component 3 to these two sampled subsets using an unpaired t test separately for each participant. Example food and non-food stimuli from this selectively sampled distribution for one subject are shown in Figure 3D along with the distribution of the selection measure (object-color probability) for food and non-food stimuli.

While the food images from this analysis visually appeared to be cool-colored, this sampling procedure, however, could result in images where the 'food' itself is warm-colored since we are computing the mean object-color probability across the entire image. We thus conducted a subsequent analysis where we sampled 50 food stimuli such that the mean object-color probability over just the food pixels (as defined using the food segmentation masks obtained from MS-COCO annotations[26]) was at the lower end of this selection measure. This yielded images where the food itself was cool-colored. We repeated the statistical analysis by comparing the responses of Component 3 to these two sampled subsets and again found that the Component 3 responds much more strongly to food than non-food stimuli (Figure S3B). This strongly suggests that the food-selectivity of Component 3 overrides any selectivity for warmer colors.

Analysis on diverse subsets of food and non-food stimuli

We selected diverse subsets of food and non-food stimuli that maximally span the representational space of different layers of a pre-trained DNN, as described above. We then compared the responses of Component 3 to these two stimulus subsets using an unpaired t test, separately for each layer and each participant. This helps us address whether food selectivity is driven by only certain kinds of food images, which would indicate that it is not 'food' selectivity per se but rather a more restricted notion that applies to only specific instances of food; or whether the selectivity even persists under conditions of wide visual variability within food and within non-food images, which would in turn indicate that it is indeed 'food' selectivity construed more broadly.

Measuring the relationship between affective features and responses of Component 3

On the subset of 100 images for which valence and arousal ratings were available (from the 'nsdmeadows' dataset), we computed the correlation between the responses of Component 3 (averaged across subjects) and subject-averaged valence and subject-averaged arousal ratings separately. For fair comparison, we also report the correlation between food salience ratings and Component 3 responses on this small subset. The statistical significance of these correlations is assessed by computing the p value of the obtained sample correlation coefficient for the null hypothesis of uncorrelation under the assumptions of a bivariate normal distribution.

Analysis on curated Natural Scenes Dataset

To test whether a high proportion of exemplars of any category in the dataset might be sufficient for a component to emerge that responds selectively to that category, we sampled a subset of images with a fixed number of examples from each of 9 categories, as described above. The question was whether we'd get equally selective components for other categories that are in the same proportion as faces and food, which might suggest that the food-selective component could arise as an artifact of the data bias in NSD. These categories are also more visually homogeneous than food (e.g. airplane). We quantified the within-category visual similarity of images by computing mean pairwise correlations between the corresponding image features in the last convolutional layer of a pre-trained CNN (layer conv5 of AlexNet trained on image categorization using ImageNet). Distances computed in the feature space of trained CNNs (versus image space) are known to correspond well to perceptual image similarity measures, and are widely used as "perceptual distance" metrics[56,57]; this metric is further also well-suited to capture similarities in mid-level visual features like texture; thus, the average pairwise image distance metric computed in this deep visual representational space for each category is likely to capture the perceptual homogeneity of that category (at least, as represented in the NSD).

We repeated the Bayesian NMF analysis on this curated dataset. On this subset, the BIC criterion suggested 7 instead of 20 components in each participant. We computed the selectivity of resulting components for each of the 9 categories using two indices: (i) Correlation (Pearson's R), where we computed the correlation of component responses to the curated stimuli with a binary vector indicating whether the category was present/absent in the image over all stimuli and (ii) t-statistic, comparing the mean responses of the component to the category in question, versus all other stimuli. For each category, we then computed the maximum selectivity value based on either of the above indices over all 7 components, as reported in Figure 5.

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

The top 2 components (based on their highest correlation with any of these category labels) in each participant were still faces and food respectively. And the highest correlation of each remaining category with all the components was substantially lower than the selectivity of the top 2 components for faces and food, respectively. This control analysis indicated that data bias (either a large number of food images in NSD or some form of visual homogeneity among the food images within NSD) cannot explain the existence of the food-selective component.

Analysis on the independent BOLD5000 dataset

We assessed whether food-selective responses can also be identified in other independent datasets beyond NSD. To test this, we analyzed data from all 4 subjects in the BOLD5000 dataset. Importantly, the shared set of 1,000 images viewed by all Phase I NSD participants were also viewed by all BOLD5000 subjects, with the exception of subject CSI4 who only viewed 594 shared images. We used these overlapping images to localize the food component in the ventral visual stream of subjects CSI1-4. This localization procedure relies on inferring the voxel weights corresponding to Component 3 (the food-selective component) in new participants using the response profile of Component 3 derived from NSD over the overlapping image set. In the non-negative matrix factorization parlance, this amounts to inferring only one weight matrix (the components by voxel weight matrix), when the other matrix (the response profile matrix) is known, subject to non-negativity constraints. Here, the latter is fixed to the Component 3 responses averaged across NSD subjects for the overlapping image set (n=1,000 for CSI1-3 and n=594 for CSI4). Mathematically, given component responses to N overlapping images derived from NSD as $\mathbf{R}$ *(Nx1)*, and the data matrix $\mathbf{D}$ (N x V) containing the responses of all V voxels to these N stimuli in a BOLD5000 subject, the non-negative voxel weights $\mathbf{W}$ (Vx1) for the component can be estimated by minimizing the expression,

$$\|\boldsymbol{R} - \boldsymbol{D}\boldsymbol{W}^{\dagger}\|_2^2 \; subject\ to\ \boldsymbol{w_i} > \mathbf{0}\ for\ all\ i = 1, ..., V$$

This optimization problem is convex and the optimal voxel weights $\hat{\mathbf{W}}$ can be derived following the standard routine for solving the non-negative least squares problem based on the active set algorithm.[58] With these inferred voxel weights, we can then estimate the component responses to novel stimuli unique to the BOLD5000 dataset ($\mathbf{D_U}$) as follows,

$$\boldsymbol{R_U} = \boldsymbol{D_U}\widehat{W}^{\dagger}$$

We restrict our focus to stimuli that had food/no-food annotations in the BOLD5000 dataset, namely, the MS-COCO images. We excluded all MS-COCO images that were viewed by any of the NSD participants from this analysis. We then computed the food-selectivity of estimated component responses to these stimuli as the correlation (Pearson's R) between component responses and a binary vector indicating whether the image contained food or not. These images were further rank-ordered by their response magnitude and colored by food labels for ease of response visualization. We further performed a control analysis by running the component localizer in other areas of the visual cortex, including early visual areas as all intermediate and high-level lateral and parietal areas, and computing the food-selectivity of the estimated component in each case. These ROIs, including the ventral visual stream ROI as used above, were defined by co-registering the *streams atlas* from NSD to the BOLD5000 anatomical space.

Quantification and statistical analysis on all component voxel weights

Agreement with functional localizer statistics

Once the voxel weights are projected back into anatomical coordinates (in the native space of each NSD participant), we can also compute the quantitative agreement between these voxel weights and the voxel-level selectivity for different categories as estimated with the independent functional localizer runs in NSD (fLOC). We computed the correlation between the voxel weights of each component against the voxel-wise t-statistic of the component's preferred category as obtained with the fLOC experiments by contrasting responses to each category against all other stimuli. For e.g. the face component voxel weights were correlated against the t-value contrasts for responses to the domain of faces over responses to all other stimuli. Note that food was not defined as a domain in the NSD fLOC experiment, since a selectivity for food in the visual cortex had never been described before; thus, we cannot perform a similar analysis for Component 3.

Anatomical similarity between saturation-responsive visual cortex and Component 3

The anatomical distribution of Component 3 appeared to overlap with previously studied color-biased regions.[59] To quantify the similarity between the anatomy of saturation-responsive regions and Component 3, we conducted a subsequent analysis. We first extracted non-food images (food salience rating of zero) from the shared set of 515 images viewed by all participants. We next computed the correlation between the saturation of all non-food stimuli (n=356) and the responses of all VVC voxels to the corresponding stimuli in order to construct a saturation-responsive voxel weight map per participant. Relationship between food selectivity and saturation-responsiveness in the ventral visual pathway is finally assessed by correlating this saturation-responsive weight map with the voxel weight map for each of the 5 components. These correlations were transformed to z-scores using Fisher's z-transformation for statistical comparisons.

Quantifying the spread of voxel weights per component

We further characterized the distribution of voxel weights for each component using quantitative measures of sparseness and statistical measures of skewness and kurtosis.

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

(a) Skewness of the voxel weight distribution w_c for each component c is computed using the Fisher-Pearson coefficient of skewness, calculated as:

Skewness $(\boldsymbol{w_c}) = \frac{m_3}{m_2^{3/2}}$, where m_2 and m_3 are respectively the second and third sample central moments of the voxel weights $\boldsymbol{w_c}$ for each component c. The rth sample moment m_r are computed using the standard formula as,

$$m_r = \frac{1}{N}\sum_{i=1}^{N}(w_{c,i} - \overline{w}_c)^r,$$

where $w_{c,i}$ is the voxel weight for component c in voxel i and $\overline{w}_c$ is the mean component weight across all N voxels. This measure is computed separately for the voxel weights per component and per participant where different participants have differing numbers of voxels (N∼6,500-9,000).

For a gaussian distribution (perfect symmetry), the skewness is zero; positive values indicate a rightward skew with more voxels that have higher weights on the component whereas negative values point towards a leftward skew.

(b) Sparseness in the voxel weights of each component, $\boldsymbol{w_c}$ with N voxels is computed using the definition of Hoyer[21] as,

$$\text{Sparseness}\ (\boldsymbol{w}_c) = \frac{\sqrt{N} - \frac{\sum_{i=1}^{N}|w_{c,i}|}{\sqrt{\sum_{i=1}^{N}w_{c,i}^2}}}{\sqrt{N} - 1},$$

Sparseness is 1 when only a single voxel has a non-zero weight on the component and is zero when all voxel weights are equal (non-sparse distribution). Values between 0 and 1 indicate intermediate levels of sparsity, interpolating smoothly between the two extremes.

(c) (Excess) Kurtosis of the voxel weight distribution for each component c is computed following Fisher's definition, as the ratio of the fourth sample central moment of the voxel weights $\boldsymbol{w_c}$ and their second central moment squared,

$$\text{Kurtosis}\ (\boldsymbol{w}_c) = \frac{m_4}{m_2^2} - 3,$$

Here, 3 is subtracted to provide a simple comparison to the Gaussian distribution which yields a kurtosis of zero under the above definition. Higher values (above 0) indicate a super-gaussian or heavy-tailed distribution indicative of sparsity.

Quantifying the inter-subject heterogeneity of voxel weights per component

We transformed the voxel weight maps from the native space of each participant to a common anatomical space (MNI 1mm) in order to measure inter-subject alignment in the anatomy of each component. For each component, this alignment was measured using correlation (Pearson's R) between the co-registered weight map of each participant and the average voxel weight map for that component across the other 7 participants (averaged across all 8 folds).

Food contrast experiment

We selected subsets of 49 'food' and 49 'non-food' images from the shared set of 515 images, as described above. We computed the food contrast on each participant individually using the conventional t-statistic comparing responses to food v/s non-food stimuli in this stimulus set, yielding values that quantify how significantly higher the response is to food stimuli compared to non-food stimuli. We did not restrict this contrast analysis to the ventral visual stream and performed this comparison on all voxels responsive to the NSD experiment (using the 'nsdgeneral' atlas released with the NSD). We then registered the contrast of each subject to the MNI space (1mm) and computed the average contrast map across all 8 NSD participants (shown in Figure 7D) for comparison with the Component 3 voxel weight map.

Encoding model of the inferred components

We used a CLIP-ResNet50[15] convolutional neural network (CNN) model to predict the response of the inferred components from the NMF analysis. The encoding model was designed to map the features from a given layer of the CNN model to the inferred responses from the component analyses (see Ratan Murty et al.[14] for more details). Importantly, we fixed all the hyper-parameters of the model based on the data from Phase 1 subjects. Specifically, we fixed the model layer (block4-1-conv2). The model features corresponding to the images used in the experiment were extracted for this layer. Next, we mapped the extracted features to the inferred component responses of Phase 2 subjects via a ten-fold regularized ridge-regression (the ridge parameter fixed at 0.01). Even though the model was trained on data from Phase 2 subjects, it was evaluated on data from Phase 1 subjects (thus cross-validating on both subjects and images). The model prediction accuracy was calculated as the Pearson correlation between the predicted response of the model (over folds) and the observed response. Our CLIP-ResNet50 encoding model is image-computable and can be used to predict the

Please cite this article in press as: Khosla et al., A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.08.009

observed responses for images not included in the NSD. We obtained predictions for: 1) The large publicly available ImageNet dataset which has diverse stimuli from 1000 stimulus categories (N = 1,281,167 images). 2) Black and white versions of the same 1.2M images as in 1, 3) color and grayscale versions of the texture-matched Downing pairs. These images were previously used to test and reject the food selectivity hypothesis in the brain (P. Downing and N. Kanwisher, 1999, Cogn. Neurosci. Soc., poster). Predictions were obtained for both color and grayscale versions of these images. (Figure 4). 4) Handpicked images that were matched across a number of stimulus features. See Figure 4 for examples. Predictions were obtained for both color and grayscale versions of these images (Figure 4).