Feature-based encoding of face identity by single neurons in the human medial temporal lobe

Runnan Cao[1], Jinge Wang[2], Chujun Lin[3], Ueli Rutishauser[4], Alexander Todorov[5], Xin Li[2], Nicholas Brandmeir[6,7], and Shuo Wang[1,7]

[1] Department of Chemical and Biomedical Engineering, West Virginia University, Morgantown, WV 26506, USA

[2] Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA

[3] Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA

[4] Departments of Neurosurgery and Neurology, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA

[5] Department of Psychology and Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA

[6] Department of Neurosurgery, West Virginia University, Morgantown, WV 26506, USA

[7] Rockefeller Neurosciences Institute, West Virginia University, Morgantown, WV 26506, USA

Corresponding authors:

Runnan Cao (runnan.cao@mail.wvu.edu)

Shuo Wang (wangshuo45@gmail.com)

Abstract

Neurons in the human medial temporal lobe (MTL) that are selective for the identity of specific people are classically thought to encode identity invariant to visual features. However, it remains largely unknown how visual information from higher visual cortex is translated into a semantic representation of an individual person. Here, we show that some MTL neurons are selective to multiple different face identities on the basis of shared features that form clusters in the representation of a deep neural network trained to recognize faces. Contrary to prevailing views, we find that these neurons represent an individual's face with feature-based encoding, rather than through association with concepts. The response of feature neurons did not depend on face identity nor face familiarity, and the region of feature space to which they are tuned predicted their response to new face stimuli. Our results provide critical evidence bridging the perception-driven representation of facial features in the higher visual cortex and the memory-driven representation of semantics in the MTL, which may form the basis for declarative memory.

Keywords: Human single-neuron recordings, Medial temporal lobe, Face, Deep neural network, Identity neuron, Feature coding

Main Text

How the brain encodes different face identities is one of the most fundamental and intriguing questions in neuroscience. There are two extreme hypotheses. The feature-based model posits that face representations are encoded over a broad and distributed population of neurons (1-4). Under this model, recognizing a particular individual requires access to many neurons, with each neuron responding to many different faces that share specific visual features such as shape and skin texture (e.g., (5) and (6)). Conclusive evidence for feature-based coding, in particular axis-based feature coding (i.e., neurons parametrically correlate with facial features along specific axes in face space), has recently been revealed in the non-human primate inferotemporal cortex (IT) (7-10). In contrast, on the other extreme, the exemplar-based model posits that explicit facial representations in the brain are formed by highly selective (sparse) but at the same time highly visually invariant neurons (11-14). Identity neurons that selectively respond to many different images showing a specific person's face embody the exemplar-based coding and are common in the human hippocampus and other parts of the medial temporal lobe (MTL) (13, 14). Recent studies have shown that the responses of identity neurons are clustered by high-level conceptual or semantic relatedness (e.g., Bill Clinton and Hillary Clinton) rather than by lower-level facial features (15, 16). Feature-based and exemplar-based models are not mutually exclusive given that both types of neurons have been observed in different brain regions; but there appears to be an abrupt transition from a distributed axis-coding model in the higher visual cortex to a sparse exemplar-based model in the MTL. The neural computations achieving this transformation remain little understood. Here, we ask the critical question of how the brain transitions from the representation of facial features processed in the higher visual cortex to the representation of identities in the MTL. We hypothesize that there are traces of feature-based encoding in the MTL and these remaining feature-based responses will enable the transformation from feature-based coding to exemplar-based coding.

To test this hypothesis, we recorded from 578 neurons in the amygdala and hippocampus (MTL areas) of 5 neurosurgical patients (16 sessions in total; **Table S1**; **Fig. S1**) while they performed a one-back task (**Fig. 1A**; accuracy = 75.7±5.28% [mean±SD across sessions]). Participants

viewed 500 natural face images of 50 celebrities (**Fig. 1**; 10 faces per identity). 490 neurons had an overall firing rate greater than 0.15Hz and we restricted our analysis to this subset of neurons, which included 242 neurons from the amygdala, 186 neurons form the anterior hippocampus, and 62 neurons from the posterior hippocampus (**Table S1**). The responses of 46/490 neurons (9.39%) differed between different face identities in a window 250-1000ms following stimulus onset (i.e., identity neurons; **Fig. 1B**; **Table S1**), consistent with prior recordings from the human MTL (13, 15, 16). Of the 46 identity neurons, 17 neurons responded to a single identity (referred to here as single-identity [SI] neurons) and the remaining 29 neurons each responded to multiple identities (referred to here as multiple-identity [MI] neurons). On average, MI neurons encoded 2.55±0.63 identities (**Fig. 1F, J**). We confirmed the results using an identity selective index (*d'* between the most- and least-preferred identities; **Fig. 1C**) and ordered responses from the most- to the least-preferred identities (**Fig. 1D**). As expected, SI neurons had a sharp decrease of response from the most-preferred identity while MI neurons showed constantly steeper changes from the most- to the least-preferred identity compared to the non-identity neurons (**Fig. 1D**). We further confirmed the results using a depth of selectivity (DOS) index (**Fig. S2A**) and single-trial population decoding (**Fig. S2B, C**), which showed that it was possible to predict the identity of the face shown.

It has been shown that some MI neurons encode *conceptually* related identities (e.g., Bill Clinton and Hillary Clinton) (15, 16) in a way that makes the response of MI neurons invariant to visual features (13, 14, 16). However, it is unknown whether MI neurons can also encode *visually* (rather than conceptually) similar identities. To answer this question, we extracted features from the images shown to the patients using a pre-trained deep neural network (DNN) VGG-16 trained to recognize faces (see **Fig. S3A, B** for DNN architecture and feature visualization). We then constructed a two-dimensional stimulus feature space using t-distributed stochastic neighbor embedding (t-SNE) feature reduction for each DNN layer (**Fig. 1G, K** and **Fig. S4**; note that quantifications below are in this t-SNE space rather than full dimensional space of the DNN; also note the pairwise distance between face examples in the full dimensional space is preserved in the t-SNE space; **Fig. S3D**). This feature space was derived solely from the input images without knowledge of neural responses and/or tuning of neurons. The feature space demonstrated an

organized structure. For example, faces of the same identity were clustered in the first fully connected (FC) layer FC6 (which is towards the top of the feature hierarchy and demonstrates clustering of identities; **Fig. S4**), where Feature Dimension 2 represented a gender dichotomy, and darker skinned faces were clustered at the bottom left corner of the feature space (**Fig. 1I, M**). Note that the DNN had no access to semantic information about the faces (e.g., gender, ethnicity, social traits), and therefore, the representation of each face in the feature space was entirely driven by visual features. Thus, faces were distributed in the feature space purely based on their visual appearance, regardless of any semantic information or conceptual association with each other.

We next projected the neuronal responses of a given neuron to each face onto this visual feature space (**Fig. 1G, K**). Strikingly, this revealed that some MI neurons were selective to different identities that were clustered in the visual feature space (**Fig. 1F-M;** see **Fig. S5** for more examples). This suggests that these neurons responded to face identities that were in fact visually similar. To formally quantify the tuning of MI neurons, we estimated a continuous spike density map in the 2D feature space from our sparse sampling (**Fig. 1H, L** upper) and used a permutation test (1000 runs; **Fig. 1H, L** lower) to identify the region that had a significantly higher spike density above chance (red outline in **Fig. 1I, M**). This region shows the part of the visual feature space to which a neuron was tuned. At the population level, we found that 13 MI neurons (44.8%) encoded all of their selective identities that were clustered in the feature space (referred to here as feature MI neurons; **Fig. 1B**; the other MI neurons encoded identities distributed in the feature space and are referred to as non-feature MI neurons). Therefore, feature MI neurons encoded visually similar identities.

In the DNN, the level of feature abstraction, and thus clustering of identities, increases from earlier layers to later layers (**Fig. S3** and **Fig. S4**). We therefore expect feature MI neurons best reflect the features in later DNN layers. Indeed, we observed feature MI neurons in DNN layers Conv5_3 (4 neurons), Pool5 (9 neurons), FC6 (11 neurons), and FC7 (11 neurons; some neurons appeared in multiple layers given the distribution of identities across layers; **Fig. S4**). The tuning region of an individual feature MI neuron covered approximately 1-3% of the 2D feature space,

with increasing coverage in higher layers (**Fig. 2A**)[1]. As expected, the distance in the face space between encoded identities decreased in higher layers when identities became more clustered (**Fig. 2B** and **Fig. S4**). As a whole, the neuronal population that we sampled covered approximately 4-10% of the feature space (**Fig. 2C**; some areas were encoded by multiple neurons). In contrast, the response of an individual SI or non-feature MI neuron covered a significantly smaller region in the feature space (**Fig. 2A**; two-tailed unpaired *t*-test: $P < 0.005$ for all comparisons) and this coverage did not increase as a function of abstraction level (**Fig. 2A**). Also, the tuning regions of the population were more distributed (**Fig. 2C**). This result was as expected because the identities (and thus the tuning regions) encoded by non-feature MI neurons were not contiguous with each other and were further apart (**Fig. 2B**). The distribution of pairwise distance between face examples within each neuron's tuning region(s) further supported this finding (**Fig. 2D**): feature MI neurons had a single large tuning region whereas non-feature MI neurons had smaller but more distributed tuning regions (shown by a bimodal distribution).

We next investigated the factors that may influence feature-based coding in identity neurons. First, previous research primarily used well-known or familiar faces to study identity neurons (13, 15, 16). It is unknown whether feature-based coding also depends on familiarity. We found that feature MI neurons encoded both familiar (i.e., patients know the name of the celebrity) and unfamiliar faces (only 54.2% of all selected identities were familiar; feature MI neurons did not differentiate familiar vs. unfamiliar selected identities; **Fig. S2D**), suggesting that face familiarity did not play an essential role for feature-based coding in the MTL (we revisit this point later with datasets consisting of all unfamiliar faces and all model synthetic faces). Second, we used a web-association metric (15) to assess whether visual similarity was distinct from conceptual similarity (**Fig. S6A**). We restricted our analysis to the faces each patient rated as familiar, but similar results were found when we included all faces in the analysis. We found that the web-association values between pairs of visually similar identities were not significantly greater than the other pairs (**Fig. 1E** left; two-tailed unpaired *t*-test: $t(22) = 0.065$, $P = 0.95$). This argues that the

[1] Note that when we calculated the tuning region, we adjusted the kernel size to be proportional to the feature dimensions such that the percentage of space coverage was not subject to the actual size of the feature space.

vicinity in the feature space was not driven by conceptual association. This was also the case for non-feature MI neurons (**Fig. 1E** right; $t(18) = 0.51$, $P = 0.62$). In fact, none of the feature spaces were correlated with conceptual associations ($P > 0.05$ for all layers; **Fig. S6B**), suggesting that the organization of the feature space was independent of the association between concepts. Therefore, it is unlikely that our findings were driven by conceptual associations. Third, we found that feature MI neurons from the same patient encoded different parts of the feature space, covering different identities (e.g., **Fig. 1J-M** vs. **Fig. S5A**), whereas feature MI neurons from different patients could encode a similar region of the feature space, covering the same identities (e.g., **Fig. 1F-I** vs. **Fig. S5B**). Furthermore, feature MI neurons were distributed across areas of the MTL (6 in the amygdala and 7 in the hippocampus) and across patients. Fourth, SI neurons and MI neurons had a similar spike sorting isolation distance (**Fig. S1H**), suggesting that MI neurons were not more likely to be multi-units. Lastly, similar results were derived if we constructed a three-dimensional feature space or used different perplexity parameters. Similar results were also derived if constructed the feature space using manifold approximation and projection (UMAP; **Fig. S7**) or principal component analysis (PCA). This suggests that our findings were robust in regard to the construction of the feature space.

Because different face examples of the same identity were not clustered in the feature space from earlier lower-level layers of the DNN (**Fig. S3C** and **Fig. S4**), we next asked whether there are neurons that code for similar visual features independent of identities. Using the same method to select identity neurons that responded to visually similar faces, we identified "feature neurons" that were tuned to a certain region of the feature space from each DNN layer (see **Fig. 3A, B** and **Fig. S8** for examples and **Fig. 3C** for a summary), regardless whether the neuron was an identity neuron. First, we found that feature neurons mostly appeared in later DNN layers where face examples started to become clustered by identities. Therefore, feature neurons from the MTL primarily encode high-level visual information related to identification rather than low-level image characteristics. The layers Conv5_3, Pool5, FC6, and FC7 contained an above-chance number of feature neurons at the population level and we restricted our analysis to these feature neurons. The number of identities (**Fig. 3D**) and face examples (**Fig. 3E**) covered by the tuning region of feature neurons indicated the size of the "receptive field" (in feature space) of these

feature neurons. The tuning region of each feature neuron covered approximately 0.5-2.5% of the feature space (**Fig. 3F**) and the neuronal population covered approximately 8-22% of the feature space (**Fig. 3G**). With increasing level of abstraction, tuning regions in later layers FC6 and FC7 were significantly larger (**Fig. 3F**; two-tailed unpaired *t*-test: $P < 0.0001$ for all comparisons), contained more face examples (**Fig. 3E**; $P < 0.05$), and had more distributed faces (**Fig. 3H**; Kolmogorov-Smirnov test: $P < 0.0001$), than earlier layers Conv5_3 and Pool5. Second, although an appreciable proportion of feature neurons were identity neurons, some feature neurons were not identity neurons (i.e., neither SI nor MI neurons; red bars in **Fig. 3C**; in particular in convolutional layer Conv5_3; see **Fig. 3A** for an example) because they covered a region in the face space containing face examples from different identities. Therefore, identity selectivity was not necessary for feature-based coding. In other words, feature neurons can respond to visually similar faces that were not from the same identity. Third, we investigated whether feature neurons were more likely to be identity neurons (i.e., either SI or MI neurons). Indeed, we found that feature neurons had a higher proportion (27/96; 28.1%) of identity neurons compared to the entire population (46/490; 9.39%; χ^2-test: $P = 2.04\times10^{-5}$; **Fig. 3I**; note that here feature neurons included those from layer Conv5_3 even though identity neurons could in principle only emerge in layers with clustering of face examples), suggesting that region-based feature tuning is a key component in identity selectivity.

The region-based feature coding we found is different from the feature coding shown in the IT of non-human primates (7-10, 17). Rather than encoding a linear combination of features, MTL neurons encoded a certain range of feature values in the face space. We previously found that some human amygdala neurons encode a linear change in facial emotions (18). Therefore, we wondered whether some MTL neurons encode a linear combination of facial features as shown in the primate IT (7, 9). To answer this question, we first used the same partial least squares (PLS) regression as in (9, 17) to identify neurons whose response could be predicted by a weighted sum of all features (i.e., feature maps; **Fig. S3B**) in the DNN (**Fig. S9A, C, G**). We also performed a linear regression of neural responses with the two dimensions of the feature space (**Fig. S4** and **Fig. S9B, D, H**; similar results were derived using face models from (7) and (19)). Using both established approaches, we did not succeed at selecting a larger than expected by

chance number of neurons that encoded a linear combination of features (**Fig. S9A, B;** see also **Fig. S9C-F** for selection within identity neurons and **Fig. S9G-J** for selection within feature neurons).

We conducted two additional experiments to validate region-based feature tuning using different stimuli and explored whether such feature coding could be generalized to unfamiliar faces. In the first new experiment, we recorded from 423 neurons in the same 5 patients (18 sessions; firing rate > 0.15Hz) using face stimuli from the FBI Twins dataset (**Fig. 4A**), which were all novel to our patients. We applied the same DNN to extract features and construct feature spaces. We again found region-based feature coding by single neurons in this experiment (see **Fig. 4A** and **Fig. S10A** for examples and **Fig. 4B-E** for group results). This suggests that feature coding by neurons in the MTL did not depend on the faces being familiar to the participants (feature coding was evident even if we restricted our analyses to the very first exposure of the faces, when they were novel). Consistent with the feature tuning derived using the above CelebA stimuli, feature neurons derived using the FBI stimuli had larger tuning regions (**Fig. 4C**; $P < 0.001$) and had more distributed faces (**Fig. 4D**; $P < 0.0001$) in layers FC6 and FC7 compared to layer Pool5. Notably, we also recorded the response of a subset of the same 330 neurons using the CelebA stimuli and we were thus able to directly investigate the generalizability of feature tuning between these two tasks. In the common feature space for the CelebA and FBI stimuli, the tuning region of 14 CelebA feature neurons overlapped with identities from the FBI stimuli (**Fig. 4F, G** and **Fig. S10C-E**). We found that FBI stimuli in the CelebA feature neurons' tuning regions elicited a significantly greater response compared to the other FBI stimuli that were not inside the CelebA feature neurons' tuning regions (paired *t*-test: $t(13) = 2.39$, $P = 0.016$; see **Fig. 4F, G** for examples and **Fig. 4H** for group results; see also **Fig. S10F** for a breakdown of each layer). This shows that region-based feature tuning generalized between different image sets as well as to novel stimuli never seen by the participant before. In the second new experiment, we recorded from a separate population of 287 neurons (13 sessions from 4 patients; firing rate > 0.15Hz) with FaceGen model faces (**Fig. 4I-K**) (19), which contained only feature information but no real identity information. Although the feature space was constructed using parameters (i.e., features) used to synthesize the faces rather than DNN features, again, we found region-based feature

tuning (25 neurons; above-chance compared to our simulations; each neuron covered 1.43% ±0.41% [mean±SD] of the feature space; see **Fig. 4I** and **Fig. S10B** for examples and **Fig. 4J, K** for group results). Together, our additional experiments showed that region-based feature tuning could be generalized to new and unfamiliar face stimuli.

In conclusion, our results reveal that the response of identity neurons in the human MTL can encode identities that are related visually rather than conceptually. We further identified feature neurons in the MTL that exhibited region-based feature coding. We showed that feature neurons were not dependent on identity selectivity nor face familiarity, and their tuning regions could be validated by new face stimuli.

The MTL is a few synapses downstream of the face-selective regions in the higher visual cortex, where feature-based coding is evident (7-10, 17)[2]. Notably, however, in the MTL so far only exemplar-based coding has been demonstrated (13, 14). This raises the question of how the brain transitions from a perception-driven representation of features in the higher visual cortex to a memory-driven representation of semantics in the MTL. Our results provide a possible mechanism by showing that MTL neurons encode a region in the high-level feature space and are selective to identities that fall in this region. The existence of a perception-driven representation of features in the MTL, importantly transitioned from axis-based to region-based, will translate visual information into an exemplar-based code. This mechanism can provide direct input to representation of semantics in the MTL, which is the basis for declarative memory (20). Therefore, our findings bridge the two extreme hypotheses by illustrating region-based feature coding in the MTL, which may form the basis for feature-invariant exemplar-based coding and semantic memory.

Neurons in the human MTL have been shown to demonstrate prominent categorical responses to visual objects (i.e., visual selectivity) (21) and facial expressions of emotions (i.e., emotion selectivity) (22, 23). Region-based feature coding may also provide an account for visual and emotion selectivity: objects or emotions falling within the coding region of a neuron may elicit

[2] It is worth noting that feature-based coding is shown in the IT of non-human primates but not yet in humans at the single-neuron level.

an elevated response. A future direction will be to construct the feature space for objects in general (e.g., using the convolutional neural network AlexNet) and investigate region-based feature coding in this feature space. Furthermore, previous research on identity neurons primarily used familiar faces (13, 15, 16). In the present study, we also found that region-based feature coding of face identity was independent of face familiarity, similar to feature coding by primate IT neurons that even encode computer generated faces (7, 9). Future studies will be needed to understand the role of memory (13, 15, 16) and attention (24) in MTL's feature coding. Lastly, it is worth noting that in contrast to the traditional axis-based face spaces where axes of the space and coordinates of face examples are fixed (7, 25), the feature space constructed by t-SNE in the present study varies as a function of the set of input stimuli because it models the similarity between all input stimuli. Therefore, our observed feature neurons in the human MTL may demonstrate a form of similarity-based or manifold-based coding (i.e., finding meaningful low-dimensional structures hidden in the high-dimensional observations using nonlinear dimensionality reduction) (26, 27), which may in turn contribute to the MTL's critical role in face recognition, classification, and memory.

Rapid advances in computer vision and development of DNNs have provided an unprecedented opportunity to help us understand the functional architecture of the brain (8, 17, 28, 29). Our present study reiterates the advantages of using DNNs to study neural encoding for face identity: by extracting features from complex natural face images using DNNs and projecting them onto the feature space constructed by DNN feature reduction, we revealed a novel face code in the human MTL that neurons encode visually similar identities.

Acknowledgements

We thank all patients for their participation, staff from WVU Ruby Memorial Hospital for support with patient testing, Minglei Yin for help on analysis, Jeremy Dawson for contributing the FBI Twins dataset, and Ralph Adolphs, Carlos Ponce, and Paula Webster for discussion and valuable comments. This research was supported by an NSF CAREER Award (1945230), ORAU

Ralph E. Powe Junior Faculty Enhancement Award, West Virginia University (WVU), WVU PSCoR Program, and the Dana Foundation (to S.W.), and an NSF Grant (OAC-1839909) and the WV Higher Education Policy Commission Grant (HEPC.dsr.18.5) (to X.L.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

R.C., A.T., X.L., and S.W. designed research. R.C. and S.W. performed experiments. N.B. performed surgery. R.C., J.W., C.L., X.L. and S.W. analyzed data. R.C, U.R., A.T., X.L., and S.W. wrote the paper. All authors discussed the results and contributed toward the manuscript.

Competing Interests Statement

The authors declare no conflict of interest.

Figure Legends

Fig. 1. Feature-based neuronal coding of face identities. **(A)** Task. We employed a one-back task, in which patients responded whenever an identical famous face was repeated. Each face was presented for 1s, followed by a jittered inter-stimulus-interval (ISI) of 0.5 to 0.75s. **(B)** Percentage of single-identity (SI) and multiple-identity (MI) neurons in the entire neuronal population. Stacked bar shows MI neurons that encoded visually similar identities (i.e., demonstrating feature-based coding; red) or not (blue). **(C)** Identity selectivity index. Both SI neurons and MI neurons had a significantly higher identity selectivity index than non-identity neurons. Error bars denote ±SEM across neurons. Asterisks indicate a significant difference using two-tailed unpaired *t*-test. ****: $P < 0.0001$. **(D)** Ordered average responses from the most- to the least-preferred identity. Non-identity neurons are shown for comparison purposes. Responses were normalized by the response to the most-preferred identity. Shaded areas denote ±SEM across neurons. The top bars indicate significant differences between SI/MI and non-identity neurons (two-tailed unpaired *t*-test, $P < 0.05$, corrected by FDR for $Q < 0.05$). **(E)** Web-association score for MI neurons. For each neuron, we calculated a mean association score between the pairs of stimuli that the neuron was selective to (S-S), and between the pairs of stimuli where the neuron was selective to one of them but not selective (NS) to the other (S-NS). Error bars denote ±SEM across neurons. Left: MI neurons that encoded visually similar identities (i.e., with feature-based coding). Right: MI neurons that did not show feature-based coding. Neither feature MI neurons nor non-feature MI neurons encoded conceptually related identities. **(F-M)** Two example neurons that encoded visually similar identities. **(F, J)** Neuronal responses to 500 faces (50 identities). Trials are aligned to face stimulus onset (gray line) and are grouped by individual identity. **(G, K)** Projection of the firing rate onto the feature space. Each color represents a different identity (names shown in the legend). The size of the dot indicates the firing rate. **(H, L)** Estimate of the spike density in the feature space. By comparing observed (upper) vs. permuted (lower) responses, we could identify a region where the observed neuronal response was significantly higher in the feature space. This region was defined as the tuning region of a neuron. **(I, M)** The tuning region of the neuron in the feature space (delineated by the red outlines).

Fig. 2. Summary of feature tuning for identity neurons. **(A)** Percentage of feature space covered by tuning regions of identity neurons. Note that here we did not apply the threshold for minimal cluster size for SI and non-feature MI neurons in order to compare between identity neurons, but we still used FDR when we identified clusters. **(B)** Normalized distance between MI neuron's selective identities in the feature space. To be comparable for different layers, Euclidean distance was normalized by the maximum distance (i.e., diagonal line) of the feature space. Error bars denote ±SEM across neurons. Asterisks indicate a significant difference between feature MI neurons and non-feature MI neurons using two-tailed unpaired *t*-test. *: $P < 0.05$, **: $P < 0.01$, and ***: $P < 0.001$. **(C)** The aggregated tuning regions of the neuronal population. Color bars show the counts of overlap between individual tuning regions. Numbers in the density map show the percentage of feature space covered by the tuning regions of the neuronal population. **(D)** Distribution of pairwise distance between face examples in each neuron's tuning region(s). Euclidean distance was normalized by the maximum distance of the feature space.

Fig. 3. Characterization of feature neurons. **(A, B)** Two example feature neurons that encoded visually similar faces. Legend conventions as in **Fig. 1**. **(C)** The number of feature neurons identified from each DNN layer. Blue: feature neurons that were also identity neurons. **(D)** The number of identities encoded by feature neurons. **(E)** The number of face examples encoded by feature neurons (i.e., the number of faces that fell within the tuning region of a feature neuron). Error bars denote ±SEM across neurons. **(F-H)** Population summary of feature tuning. Legend conventions as in **Fig. 2**. **(I)** The number of identity neurons in the whole population (left) and among feature neurons (right). Blue: the number of identity neurons. Red: the number of non-identity feature neurons. Gray: the number of non-identity neurons.

Fig. 4. Validation and generalization of feature tuning with unfamiliar and model faces. **(A-H)** Results from the FBI Twins dataset. **(I-K)** Results from the FaceGen dataset. **(A)** An example

neuron demonstrating region-based feature coding. In FBI face spaces, similar faces were also clustered and faces from different genders were organized in different areas of the feature space. The size of the dot indicates the firing rate. The red outline delineates the tuning region of the neuron in the feature space. **(B)** The number of identified feature neurons using FBI stimuli. Only DNN layers with an above-chance number of feature neurons are shown. **(C-E)** Population summary of feature tuning. Legend conventions as in **Fig. 2**. **(F, G)** Example CelebA feature neurons showing elevated responses for FBI stimuli falling in their tuning regions. Feature spaces were constructed for combined CelebA and FBI stimuli. The size of the dot indicates the firing rate. The red outline delineates the tuning region of the neuron (identified by the CelebA stimuli). Black: face examples from the CelebA stimuli. Gray: face examples from the FBI stimuli. Magenta: FBI stimuli falling in the tuning region of the neuron. **(H)** Population results comparing neuronal response to FBI stimuli falling in vs. out of the tuning region. Each dot represents a neuron. Error bars denote ±SEM across neurons. Asterisk indicates a significant difference between In vs. Out responses using paired *t*-test ($P < 0.05$). **(I)** An example neuron demonstrating region-based feature coding. The dimensions of the feature space are the first shape and texture principal components (PCs) used to generate the stimuli. Note that face shape varied along Feature Dimension 1 and skin color varied along Feature Dimension 2. **(J, K)** Population summary of feature tuning. Legend conventions as in **Fig. 2**.

References

1. W. J. Freeman, Mass action in the nervous system. (Citeseer, 1975), vol. 2004.
2. G. E. Hinton, Distributed representations. (1984).
3. E. T. Rolls, A. Treves, M. J. Tovee, The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. *Experimental Brain Research* **114**, 149-162 (1997).
4. P. S. Churchland, T. J. Sejnowski, *The computational brain*. (MIT press, 2016).
5. M. A. Turk, A. P. Pentland, in Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (IEEE, 1991), pp. 586-591.
6. W. A. Freiwald, D. Y. Tsao, M. S. Livingstone, A face feature space in the macaque temporal lobe. *Nat Neurosci* **12**, 1187-1196 (2009).
7. L. Chang, D. Y. Tsao, The Code for Facial Identity in the Primate Brain. *Cell* **169**, 1013-1028.e1014 (2017).
8. P. Bashivan, K. Kar, J. J. DiCarlo, Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
9. C. R. Ponce *et al.*, Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell* **177**, 999-1009.e1010 (2019).
10. P. Bao, L. She, M. McGill, D. Y. Tsao, A map of object space in primate inferotemporal cortex. *Nature*, (2020).
11. H. B. Barlow, Single Units and Sensation: A Neuron Doctrine for Perceptual Psychology? *Perception* **1**, 371-394 (1972).
12. T. Valentine, A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A* **43**, 161-204 (1991).
13. R. Quian Quiroga, L. Reddy, G. Kreiman, C. Koch, I. Fried, Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102-1107 (2005).
14. R. Quian Quiroga, Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience* **13**, 587 (2012).
15. E. De Falco, M. J. Ison, I. Fried, R. Quian Quiroga, Long-term coding of personal and universal associations underlying the memory web in the human brain. *Nature Communications* **7**, 13408 (2016).
16. H. G. Rey *et al.*, Single Neuron Coding of Identity in the Human Hippocampal Formation. *Current Biology*, (2020).
17. D. L. K. Yamins *et al.*, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* **111**, 8619 (2014).
18. S. Wang *et al.*, The human amygdala parametrically encodes the intensity of specific facial emotions and their categorical ambiguity. *Nature Communications* **8**, 14821 (2017).
19. N. N. Oosterhof, A. Todorov, The functional basis of face evaluation. *Proceedings of the National Academy of Sciences* **105**, 11087-11092 (2008).

20. U. Rutishauser, Testing Models of Human Declarative Memory at the Single-Neuron Level. *Trends in Cognitive Sciences* **23**, 510-524 (2019).
21. G. Kreiman, C. Koch, I. Fried, Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci* **3**, 946-953 (2000).
22. I. Fried, K. A. MacDonald, C. L. Wilson, Single Neuron Activity in Human Hippocampus and Amygdala during Recognition of Faces and Objects. *Neuron* **18**, 753-765 (1997).
23. S. Wang *et al.*, Neurons in the human amygdala selective for perceived emotion. *Proceedings of the National Academy of Sciences* **111**, E3110-E3119 (2014).
24. S. Wang, A. N. Mamelak, R. Adolphs, U. Rutishauser, Encoding of Target Detection during Visual Search by Single Neurons in the Human Brain. *Current Biology* **28**, 2058-2069.e2054 (2018).
25. D. A. Leopold, I. V. Bondar, M. A. Giese, Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* **442**, 572-575 (2006).
26. S. T. Roweis, L. K. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **290**, 2323 (2000).
27. J. B. Tenenbaum, V. d. Silva, J. C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290**, 2319 (2000).
28. K. Grill-Spector, K. S. Weiner, J. Gomez, A. Stigliani, V. S. Natu, The functional neuroanatomy of face perception: from brain measurements to deep neural networks. *Interface Focus* **8**, 20180013 (2018).
29. S. Grossman *et al.*, Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature Communications* **10**, 4934 (2019).

Figure 1

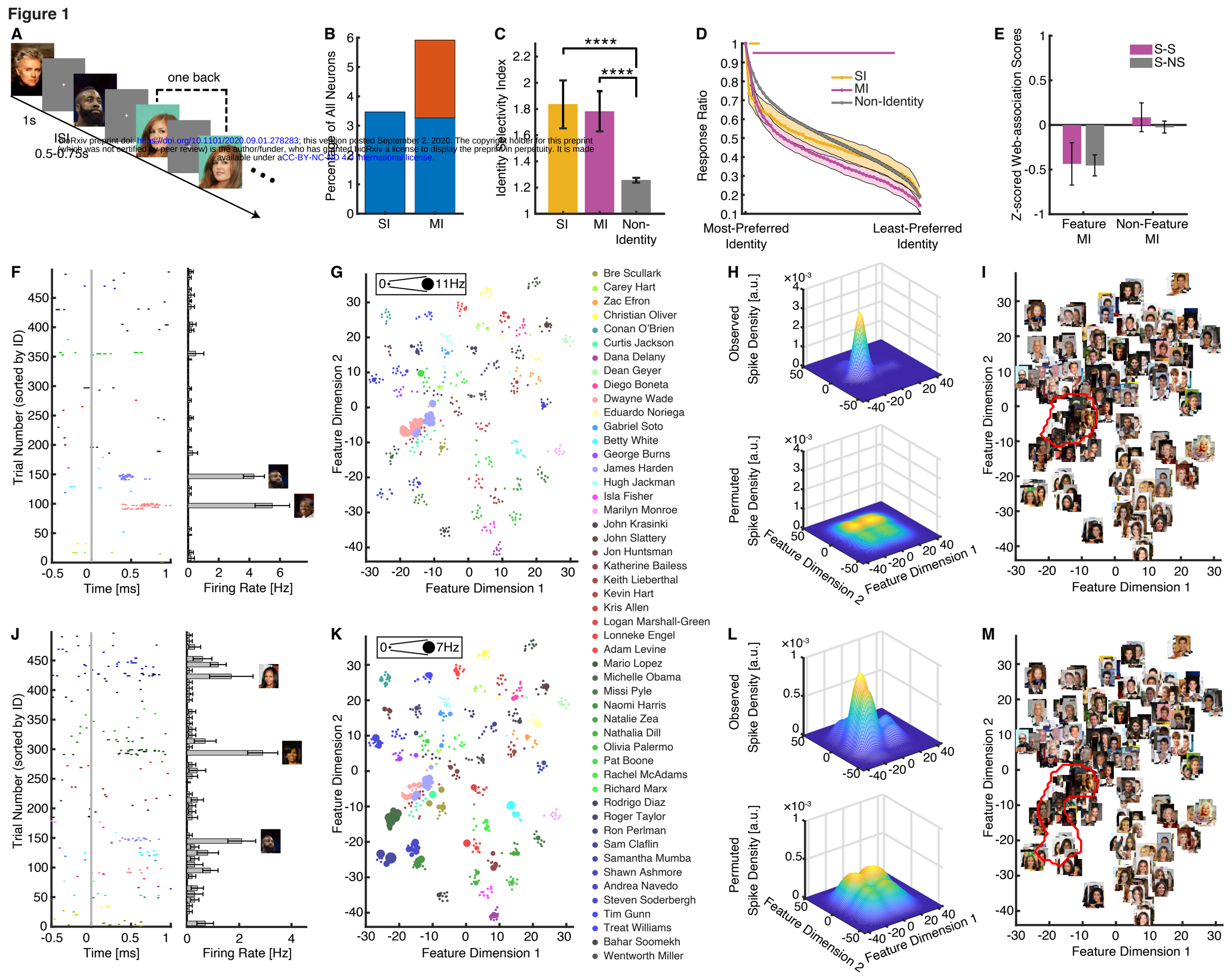

Figure 2

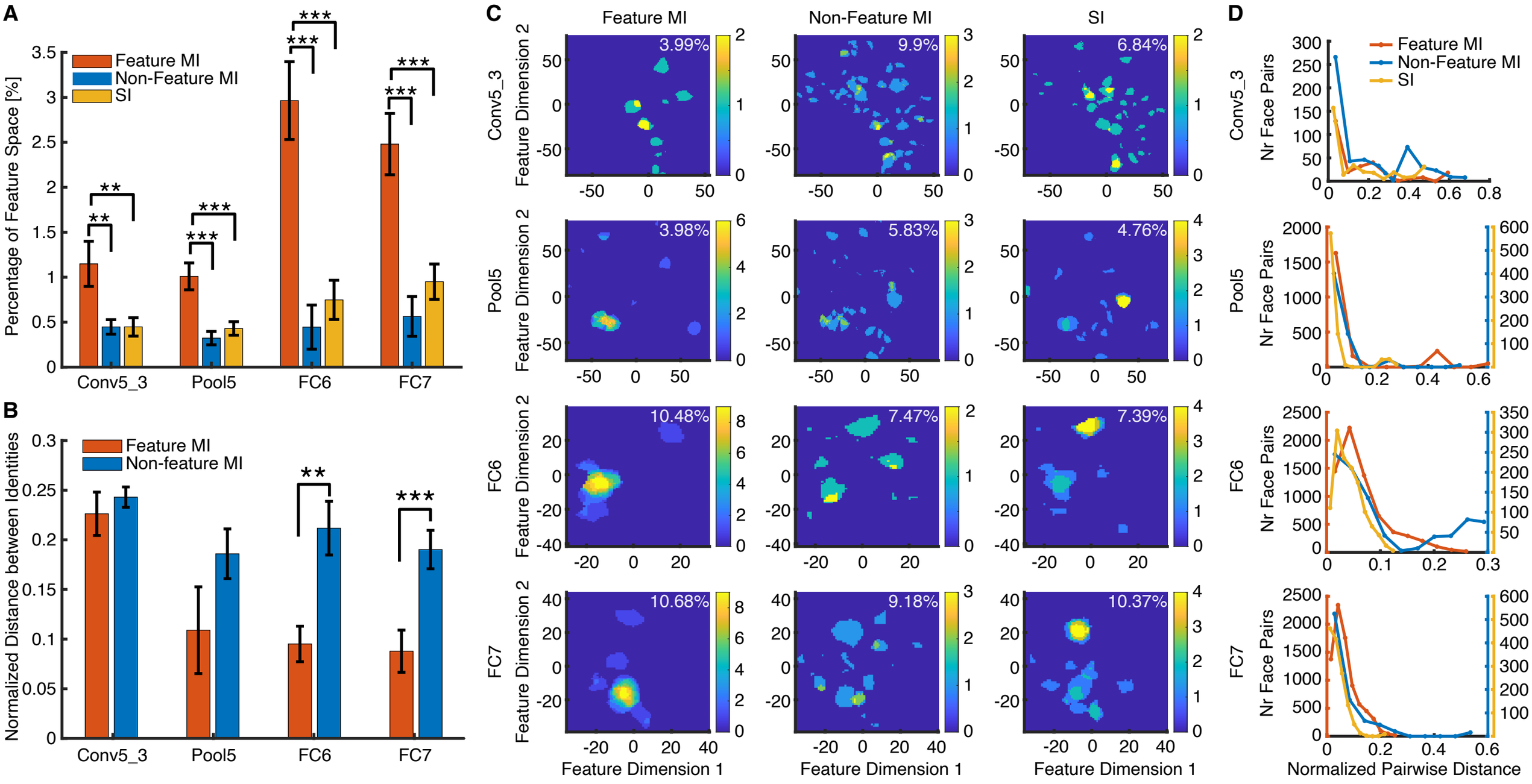
A
Feature MI
Non-Feature MI
SI
Percentage of Feature Space [%]
Conv5_3
Pool5
FC6
FC7
B
Feature MI
Non-feature MI
Normalized Distance between Identities
C
Feature MI
Non-Feature MI
SI
Feature Dimension 2
Feature Dimension 1
3.99%
9.9%
6.84%
3.98%
5.83%
4.76%
10.48%
7.47%
7.39%
10.68%
9.18%
10.37%
D
Nr Face Pairs
Normalized Pairwise Distance

Figure 3

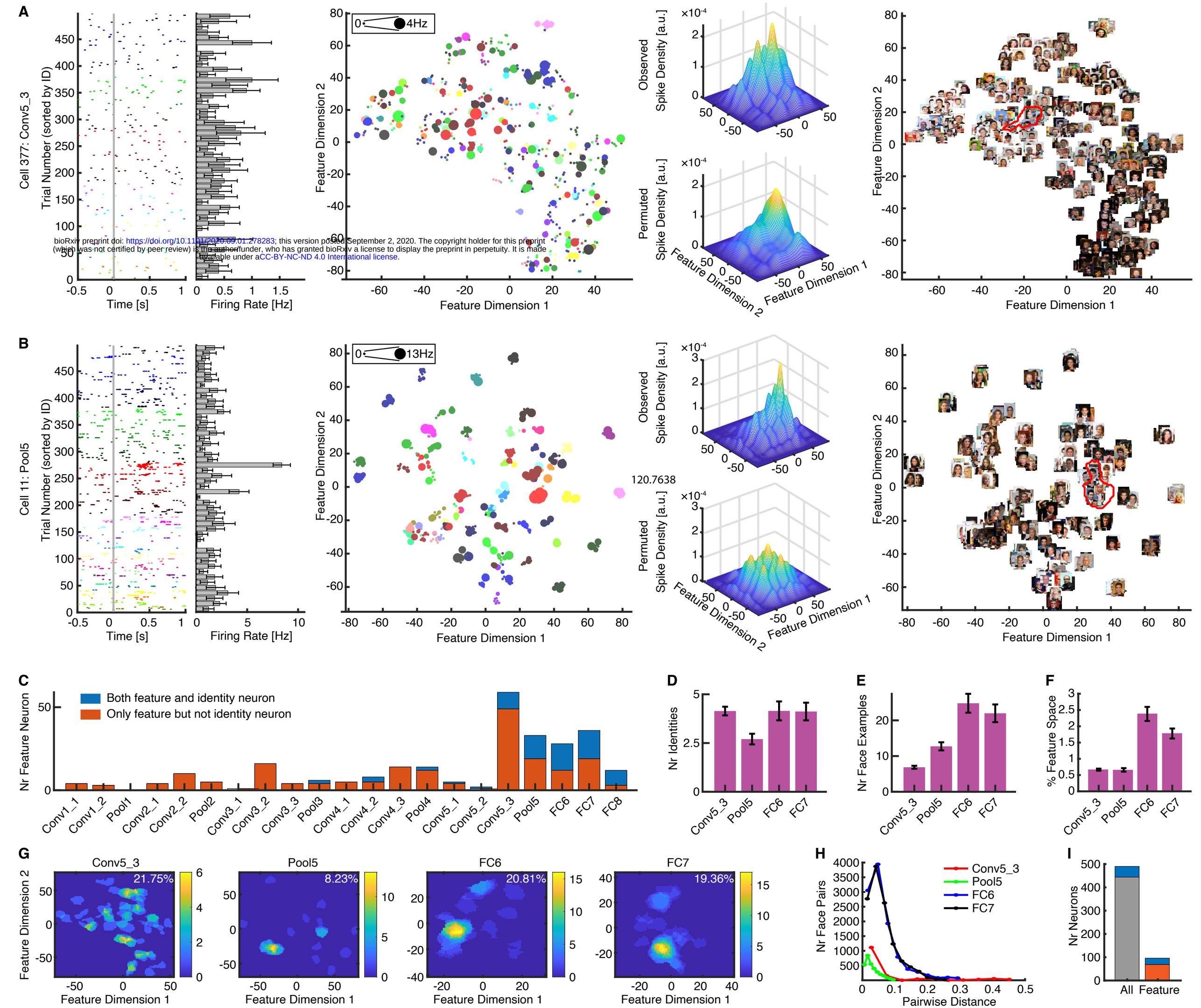

Figure 4