Neuron

Article

Predictive coding of natural images by V1 firing rates and rhythmic synchronization

Highlights

- Predictability in natural images quantified with self-supervised neural networks
- V1 firing rates decrease with predictability of high- not low-level image features
- γ-synchronization increases with predictability of low-level image features
- Late-onset β-synchronization for natural scenes with low predictability

Authors

Cem Uran, Alina Peter, Andreea Lazar, ..., Pascal Fries, Wolf Singer, Martin Vinck

Correspondence

cem.uran@esi-frankfurt.de (C.U.), martin.vinck@esi-frankfurt.de (M.V.)

In brief

Uran, Peter et al. use self-supervised neural networks to quantify stimulus predictability in natural images to investigate the context-dependence of V1 signals. Firing rates decrease with the predictability, specifically of high-level image features. By contrast, γ-synchronization increases with the predictability of low-level features and emerges for low-dimensional, strongly compressible images.

Uran et al., 2022, Neuron *110*, 1–18
April 6, 2022
https://doi.org/10.1016/j.neuron.2022.01.002

CellPress

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

Neuron

CellPress
OPEN ACCESS

Article

Predictive coding of natural images by V1 firing rates and rhythmic synchronization

Cem Uran,[1,5,6,*] Alina Peter,[1,6] Andreea Lazar,[1] William Barnes,[1,2] Johanna Klon-Lipok,[1,2] Katharine A. Shapcott,[1,3] Rasmus Roese,[1] Pascal Fries,[1,4] Wolf Singer,[1,2,3] and Martin Vinck[1,5,7,*]

[1]Ernst Strüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, 60528 Frankfurt, Germany
[2]Max Planck Institute for Brain Research, 60438 Frankfurt, Germany
[3]Frankfurt Institute for Advanced Studies, 60438 Frankfurt, Germany
[4]Donders Institute for Brain, Cognition and Behaviour, Department of Biophysics, Radboud University Nijmegen, 6525 AJ Nijmegen, the Netherlands
[5]Donders Centre for Neuroscience, Department of Neuroinformatics, Radboud University Nijmegen, 6525 AJ Nijmegen, the Netherlands
[6]These authors contributed equally
[7]Lead contact
*Correspondence: cem.uran@esi-frankfurt.de (C.U.), martin.vinck@esi-frankfurt.de (M.V.)
https://doi.org/10.1016/j.neuron.2022.01.002

SUMMARY

Predictive coding is an important candidate theory of self-supervised learning in the brain. Its central idea is that sensory responses result from comparisons between bottom-up inputs and contextual predictions, a process in which rates and synchronization may play distinct roles. We recorded from awake macaque V1 and developed a technique to quantify stimulus predictability for natural images based on self-supervised, generative neural networks. We find that neuronal firing rates were mainly modulated by the contextual predictability of higher-order image features, which correlated strongly with human perceptual similarity judgments. By contrast, V1 gamma (γ)-synchronization increased monotonically with the contextual predictability of low-level image features and emerged exclusively for larger stimuli. Consequently, γ-synchronization was induced by natural images that are highly compressible and low-dimensional. Natural stimuli with low predictability induced prominent, late-onset beta (β)-synchronization, likely reflecting cortical feedback. Our findings reveal distinct roles of synchronization and firing rates in the predictive coding of natural images.

INTRODUCTION

There are widely different theoretical accounts of stimulus processing in visual cortex. Feedforward (FF) models of vision explain neural responses based on receptive field (RF) properties that arise through FF convergence and span a circumscribed region of space in early visual areas. A central idea of these models is that neurons in visual cortex extract features that allow for solving tasks such as object classification. In support of this view, there are close similarities between the stimulus-response properties of units in the primate ventral stream and convolutional neural networks for object recognition (OR-CNNs). Other theories of vision such as predictive and efficient coding have assigned a central role to the integration of sensory evidence with contextual information. For such integration, recurrent dynamics, implemented through lateral and top-down connections, are thought to play an important role. Predictive coding postulates that a key computational goal of sensory circuits is to perform active perceptual inference based on sparse sensory data, and to encode information efficiently (Rao and Ballard, 1999; Mumford, 1992; Friston, 2010; Srinivasan et al., 1982; Von Helmholtz, 1867; Haefner et al., 2016). Predictive processing could also play a central role in self-supervised learning, which has been challenging to implement in networks trained on object classification (Bengio et al., 2012; LeCun, 2019). In contrast to object classification where labels for learning are sparse under natural conditions, contextual predictions can be continuously compared with sensory inputs. To understand the potential role of predictive coding mechanisms in visual cortex, a key question is how stimuli with weak versus strong contextual predictability are encoded and transmitted by neuronal populations.

The predictive coding model of Rao and Ballard (1999) postulates that neuronal populations signal surprising information through enhanced firing rates, whereas predictable information entails a suppression of neural activity. Consistent with this model, firing rates in visual areas such as V1 are enhanced for small compared with large natural stimuli (i.e., when lacking spatial context) (Vinje and Gallant, 2000; Coen-Cagli et al., 2015) and for artificial stimuli that have non-matching features across space (e.g., orientation or color) (Bair et al., 2003; Peter et al., 2019). Spatial context may not only determine firing rates,

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

but likely also recurrent interactions between neuronal populations, thereby inducing specific temporal patterns of correlated firing among neurons. These temporal correlations can carry additional information (compared with firing rates alone) about contextual predictability and could coordinate the interactions between neuronal groups both within and between areas. In area V1, many visual stimuli induce quasi-oscillatory states in the γ-frequency range (30–80 Hz), which are accompanied by synchronized firing over several millimeters in cortical space, called "gamma (γ)." Gamma activity is thought to result from balanced interactions between inhibitory and excitatory neurons (Onorato et al., 2020; Vinck et al., 2013; Spyropoulos et al., 2020). It has been proposed that γ-synchronized firing promotes FF processing of information and is enhanced when stimuli are salient and attended (Fries et al., 2001; Biederlack et al., 2006; Fries, 2015; König et al., 1996; Bastos et al., 2015). Building on this idea, it has been proposed that cortical γ-synchronization mediates the signaling of prediction errors and that slower rhythms such as α and β carry prediction signals that require integration on longer timescales (Bastos et al., 2020, 2012; Arnal et al., 2011; Chao et al., 2018) (but see Ferro et al., 2021). In contrast to this proposal, it has also been hypothesized that in visual cortex, γ-synchronization among spiking discharges occurs predominantly in states of high stimulus predictability (Vinck and Bosman, 2016). In this scenario, synchronization could carry complementary information to firing rates by signaling a match between prediction and evidence, which may be of equal importance to signal transmission and plasticity as error signals (Singer, 2021; Grossberg, 1987).

Contextual predictions should reflect our innate and learned priors about the statistics of the natural environment. Hence, it is critical to investigate the neural correlates of predictability using natural scenes rather than artificial stimuli (e.g., gratings, homogeneous surfaces), which are extreme outliers in our visual environment. It has remained difficult to study the distinct roles of firing rates and synchronization in a general form for natural images because it is unclear how the constructs of predictions and predictability should be operationalized and quantified. Ideally, this would rely on neural networks that learn both linear and non-linear natural scene statistics (i.e., priors) across a very large number of images in a self-supervised manner. These natural scene statistics contain low-level (pixel structure) to high-level (object information) features for which biological neurons, with encoding properties shaped by natural scene priors, could encode sensory predictions or prediction errors. This suggests that there may not be one all-encompassing measure of predictability, but that different levels of predictability should be distinguished. Here, we derived measures to assess predictability in natural images in order to investigate the contextual modulation of firing rates and synchronization in macaque V1. To this end, we developed a self-supervised deep neural network (DNN) to generate predictions of the likely structure of stimuli falling in a neuron's RF based on natural image statistics. By comparing these predictions to the actual stimuli, we obtained different measures of predictability and related them to neural activity. Our data suggest distinct roles for firing rates and synchronization in the predictive processing of natural scenes.

RESULTS

Three macaque monkeys performed a passive fixation task and viewed large (>11°) natural images. We recorded multi-unit (MU) spiking activity and local field potentials (LFPs) from 32–64 channel microelectrode arrays in area V1 (Figure 1A). RF eccentricities were 5.2 degrees of visual angle (dva) on average (range: 2.5–10.6 dva) with an average diameter of 1.44 dva (which was likely overestimated because of small eye movements). LFP and multi-unit activity (MUA) power spectra showed a typical $1/f$ trend with characteristic peaks in the γ-frequency band (30–80 Hz) (Figures S1A–S1D; Pesaran et al., 2018). These γ-peaks are known to reflect the rhythmic synchronization of synaptic and spiking activity (Onorato et al., 2020; Buzsáki and Draguhn, 2004; Pesaran et al., 2018). Their magnitude was estimated by removing the $1/f$ trend and fitting polynomials as in Peter et al. (2019) (see STAR Methods).

Quantifying predictability using deep neural networks

We developed a method to quantify the stimulus predictability of visual inputs to the RF (see STAR Methods; Figure 1B). A DNN was trained in a self-supervised way to form stimulus predictions about a small region of the image based on the rest of the image, which can be understood as a form of predictive coding. During training, stimulus predictions were improved iteratively, each time comparing the predicted RF image-patch with the ground-truth image patch. The input to the network was the image region surrounding the RF image-patch, i.e., the embedding context (224 × 224 pixels; ~5–6 dva wide). Based on this contextual input, the network generated a predicted RF image-patch (which was set to 1 dva, roughly corresponding to the size of the neuron's RFs). The network was trained using natural images that were not presented during recordings (see STAR Methods). Training was self-supervised because only data from the image itself were used for training, without human labels. As a loss function for training, we quantified perceptual similarity between the ground-truth RF image-patch and the predicted RF image-patch.

As novel input to this trained network, we then used images presented during recordings. The network predicted the content of an RF image-patch, which was centered on a given recording site's RF center, based on the image context. We then used a perceptual similarity measure in order to quantify the difference between the predicted and ground-truth patch. This perceptual similarity measure computed pixel-wise correlations, which is comparable with the well-established perceptual similarity measure structural similarity index (SSIM) (Figure S1E). The computed value, called "structural predictability," reflects the extent to which the precise pixel structure of a given stimulus falling in the V1 RF can be predicted by the spatial context. This represents a predictability measure on the image level. Different levels of predictability will be considered further below.

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

CellPress
OPEN ACCESS

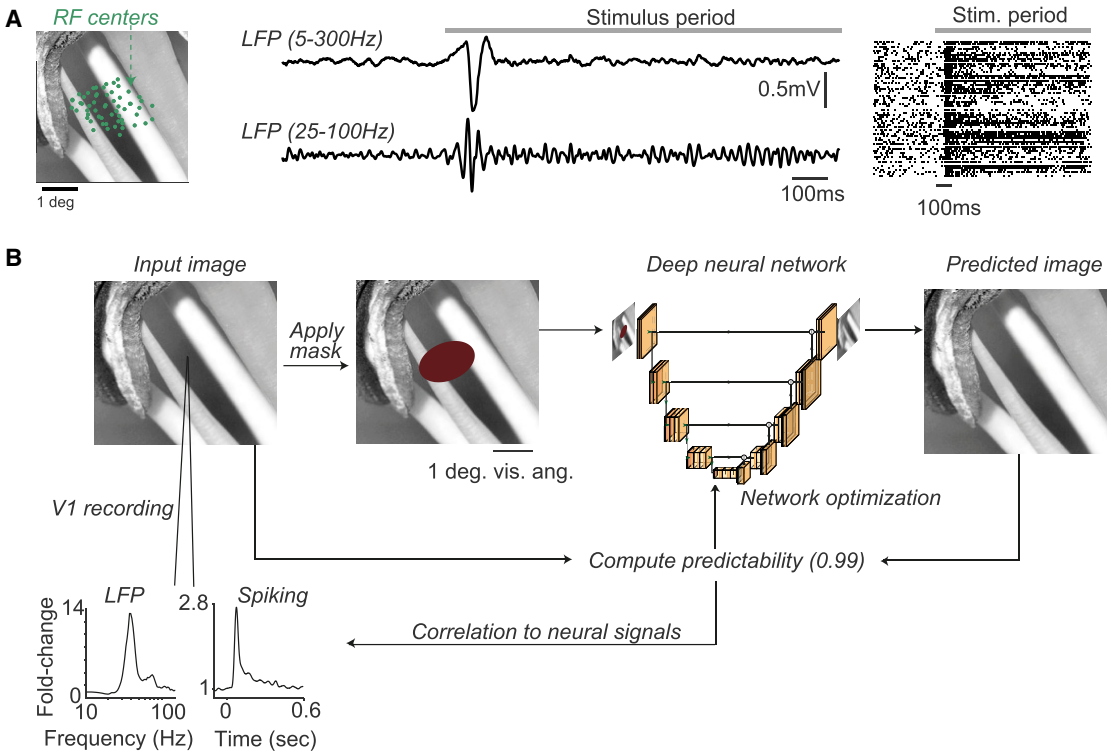

Figure 1. Recording paradigm and machine learning method to compute predictability for natural scenes

(A) Natural images were presented for 1.2 s (in a subset of sessions, for 0.6 s). (Left panel) Green dots indicate locations of RF centers of the recording array in monkey H. Image is cut out around the RF locations. (Center) Median example trace of the LFP for the image shown on the left. The 25–100 Hz filtered trace has arbitrary units. (Right) Example raster plot for MUA (spikes threshold at 3 s.d).

(B) Illustration of a deep neural network (DNN) trained to predict visual inputs into the RFs. A mask of approximately the same size as the recording site's RF is applied to an image. The image with the mask is then entered as an input to a DNN with a U-net architecture. This DNN generates (predicts) the full image, i.e., the image content behind the mask is filled in. Stimulus predictability is computed by comparing the ground-truth input image and the predicted image and then used for network optimization during the training stage. After network training, a novel set of images is presented to both the DNN and the monkeys. The predictability score is then correlated with LFP and spiking responses across images. See also Figure S1.

Predictability in natural images determines synchronization of V1 activity

We then examined the relationship between neural activity and structural predictability. LFP spectra showed major differences depending on the degree of structural predictability. For strong structural predictability, LFP power spectra showed a prominent narrow-band peak in the γ-frequency range, which was several times larger than pre-stimulus power (Figures 2A–2C, **S2A**, and **S2B**). By contrast, LFP spectra did not show γ-peaks for weak structural predictability. Across structural predictability bins, γ-synchronization increased monotonically and by ~300% (Figures 2B and 2C). This finding was consistent across the three monkeys (H, I, and A: Pearson's r = 0.94, 0.85, and 0.87), and similar findings were made for MUA spiking activity in the γ range (Figure S2C). In an additional analysis, we directly correlated LFP γ-power with structural predictability across all images (i.e., without using quantiles, see STAR Methods). There was a clear positive correlation between γ-synchronization and structural predictability (Figure 2D). Findings were consistent between early and late trials for a given stimulus in a session, and the correlation with predictability was observed already during the first presentation (Figure S2F). Furthermore, the findings were not explained by eye movements or pupil diameter (Figure S2E).

Unexpectedly, we observed that LFPs showed a prominent peak in the (high) β-frequency band (18–30 Hz) for stimuli with weak structural predictability (Figures 2A–2C). This β-peak emerged only during the late phase of the stimulus period (>500 ms) (Figures 2A and S2A). LFP β-peaks were detected in only 2/3 animals and were negatively related to structural predictability in both (monkey H: Pearson's r = −0.79; monkey I: r = −0.54; note that in the third monkey, stimulus duration was only 600 ms). MUA showed considerably less rhythmicity in the β-range for low predictability compared with the γ-rhythmicity observed for high structural predictability (Figure S2C). For the rest of the paper, we focus primarily on firing rates and γ, and report several of the main analyzes for β in the supplementary figures.

Firing rate intensity was computed separately for early (50–150 ms) and late stimulus periods (200–600 ms), by determining the instantaneous energy of the MU activity (see STAR Methods). Note that in this study, we primarily focus on the late rates (1) because surround modulation was the strongest in this period (see further below), (2) to compare γ and rates in

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

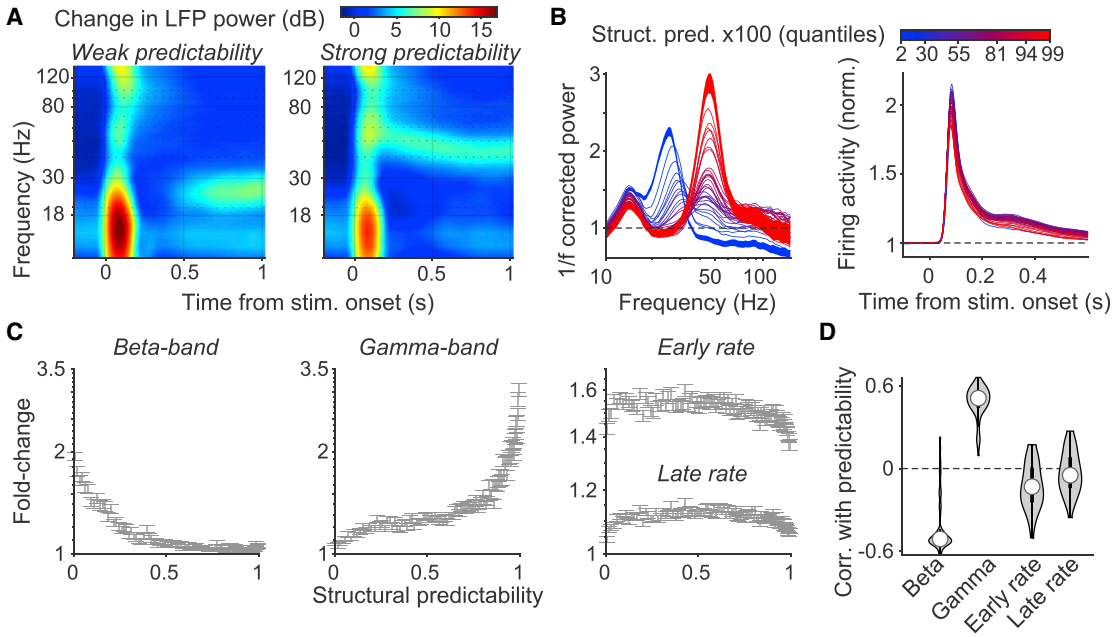

Figure 2. Distinct relationships of firing rates and neural synchronization with structural predictability

(A) Average time-frequency representations for weak and strong structural predictability.

(B) (Left) Average 1/*f*-corrected LFP power spectra for monkey H, for different levels of structural predictability. Black line indicates the pre-stimulus period. SEMs are shown only for the lowest and highest quantile of structural predictability. (Right) Multi-unit firing rates.

(C) (Left) Average (±SEM) 1/*f*-corrected β-peak amplitude versus structural predictability. Average was computed across all recording sites in the three animals (n = 72 sites). (Middle) Same for γ. (Right) Same for early (50–150 ms) and late firing rates (200–600 ms).

(D) Pearson-r correlation across recording sites (with a minimum RMS contrast of 0.1) between structural predictability and γ, β, and rate. Correlations were computed for each recording site separately, using all images presented across sessions. Correlations were significant for β, γ, and early rates (all $p < 0.001$) but not for late rates ($p = 0.11$) (t test). Absolute correlations were higher for γ and β than early and late rates ($p<0.001$ for all comparisons). Data in (B and C) are represented as mean ± SEM. See also Figure S2.

the same period, and (3) to avoid the early onset transient. Early firing rates showed weakly negative correlations with structural predictability, and late firing rates did not have significant correlations with structural predictability (Figures 2B–2D). Firing intensity was non-monotonically related to structural predictability, reaching maximal values for intermediate structural predictability (Figure 2C). These findings would thus appear to contradict the hypothesis that V1 firing rates encode sensory prediction errors (Rao and Ballard, 1999). However, as shown below, firing rates modulations are primarily determined by a different form of stimulus predictability, namely high-level stimulus predictability.

Predictability and data compression

To further understand the significance of structural predictability for visual encoding, we related structural predictability to data compression. We reasoned that natural images with predictable structure have a high degree of redundant information and should therefore be highly compressible. To compute data compression rates, we used a large image database in which we compressed each image and determined the number of bits per pixel in the compressed image (see STAR Methods). For each image, we also computed the average structural predictability across 16 image locations (examples shown in Figure 3A). Structural predictability was strongly correlated with image compressibility (Figure 3B). Thus, images with high structural predictability can also be efficiently encoded by an image compression algorithm.

Based on the results shown in Figure 2, we reasoned that γ but not firing rates should correlate strongly with compressibility. To investigate this, we computed the compressibility of the 3 × 3 degree image patches centered on the RFs and correlated this with neural activity. We found that γ showed a much stronger correlation with compressibility than firing rates (Figures 3C and S3A).

Predictability, dimensionality, and natural image statistics

Next, we wondered how neural activity relates to the dimensionality of visual inputs. We expected that when the RF input can be well predicted by the surrounding context, the image can be represented by relatively few spectral components, which can be understood as a low dimensionality. For example, a grating or a homogeneous surface can be represented by a single spectral component. To quantify dimensionality, we took the rotational average of the two-dimensional Fourier transform of the RF image-patch, sorted the spectral components by their magnitude and computed the slope of the spectrum (Figure S3Bi). Image patches with high structural predictability had significantly lower dimensionality (Figures 3D, right and **S3Bii**). Accordingly,

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

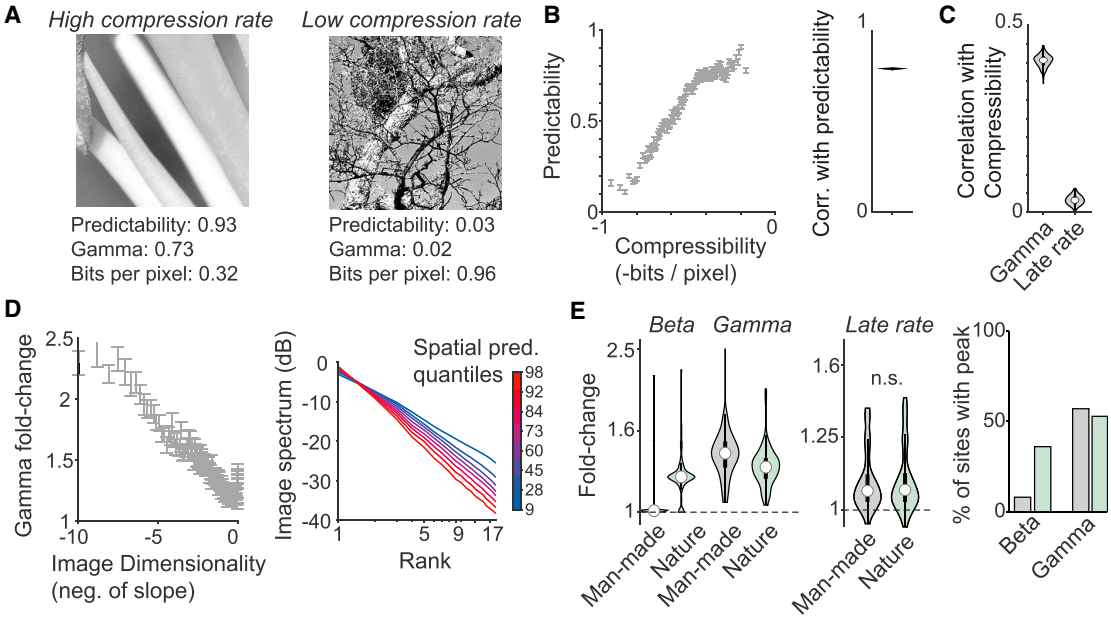

Figure 3. Synchronization reflects image compressibility and dimensionality and distinguishes natural image categories

(A) Two examples of images that have a low and high compression rate, structural predictability and γ values (as $\log_{10}$-fold change), respectively. Compression rate was measured as the number of bits/pixel for image compression.

(B) (Left) Compressibility (i.e., negative of compression rate) versus average structural predictability. (Right) Correlation between compressibility and structural predictability across images.

(C) Average correlation (across recording sites) between compressibility and γ and firing intensity across images.

(D) (Left) Images with low dimensionality had strong γ synchronization (r across quantiles = -0.9, $p<0.001$). Dimensionality was determined from the slope of the image spectrum. (Right) Average magnitude of spectral image components versus structural predictability (Pearson's r = -0.91, $p<0.001$).

(E) Fold-changes in neural activity for images with man-made content or nature content in the RF. Comparison was significant for β ($p<0.001$) and γ ($p< 0.001$), but not for early and late rates ($p = 0.5$ and $p = 0.9$, t test). (Right) Percentage of sites with a detectable β or a γ peak, across all randomly selected images. Data in (B and D) (left) are represented as mean ± SEM. See also Figure S3.

γ-synchronization was maximal for images with low dimensionality and decreased monotonically as dimensionality increased (Figure 3D, left).

We further asked if neural activity could distinguish between different image categories (Torralba and Oliva, 2003). Images with man-made objects often contain predictable structure, whereas less predictable structure is common for images of nature (Figures S3C and S3D). Furthermore, predictable structure is associated with the presence of object boundaries covering the RFs. Firing rate intensity did not differ between nature and man-made categories and did not depend on the presence of an object boundary (Figures 3E and S3E). β-synchronization was stronger for images of nature, whereas γ-synchronization was stronger for images with man-made content and object boundaries in the RFs (Figures 3E and S3E). Accordingly, stimuli such as gratings, straight and curved bars, and edges of filled contours generated strong γ (Figures S3G–S3J). Structural predictability showed high variability in natural images (Figure S3F). There was also substantial variability in γ and β across images: γ peaks were detected only for about 50% of sites, and β peaks were found almost exclusively for the nature category (Figure 3E).

Finally, we analyzed whether our findings extended to color images because color influences both γ and firing rates (Shirhatti and Ray, 2018; Peter et al., 2019; Wachtler et al., 2003; Rols et al., 2001). In separate sessions, we presented images both in grayscale and their original color. Across images, γ-and β-synchronizations showed a strong correlation between grayscale and color images (Pearson's r = 0.81 and 0.7, respectively, $p<0.001$ for both).

Relationship with salience

Because we found γ-synchronization to be positively correlated with contextual predictability, and it is known that contextual predictability is an important determinant of image salience (Li, 2002), we wondered whether γ-synchronization is correlated with image salience. To investigate this, we extracted salience maps from each image using a state-of-the-art DNN (see STAR Methods). These networks operationalize salience by predicting in which areas of an image the subjects prefer to direct their gaze. Indeed, we found that image salience correlated negatively with γ-synchronization (Figure S4I), consistent with the relatively strong negative correlation between predictability and image salience (see Figure S6D). The negative correlation of predictability with salience had a substantially greater magnitude than the positive correlation of salience with luminance contrast (Figure S6D). This finding is in line with the lack of a straightforward relationship between luminance contrast and salience (Einhäuser and König, 2003) and the importance of contextual predictability (Li, 2002).

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

Interaction of stimulus predictability with stimulus drive

In addition to predictability, natural images also show substantial variability in luminance contrast (Figure S4A). Luminance contrast influences the contextual modulation of sensory responses (Sceniak et al., 1999; Kapadia et al., 1999; Cavanaugh et al., 2002) and correlates positively with γ-synchronization for artificial grating stimuli (Henrie and Shapley, 2005; Roberts et al., 2013; Hadjipapas et al., 2015). We therefore wondered how the influence of predictability on neural activity depended on luminance contrast.

For each recording site, we computed the luminance contrast (specifically root mean square contrast) of its corresponding RF image-patch (see STAR Methods). Firing rates increased as a function of luminance contrast (Figures 4A and 4B). Luminance contrast also correlated positively with γ-synchronization (Figures 4A and 4B; for β see Figure S4B), and the γ peak frequency showed a moderate increase with luminance contrast (Figure 4A), consistent with previous work using grating stimuli (Henrie and Shapley, 2005; Ray and Maunsell, 2010; Hadjipapas et al., 2015; Roberts et al., 2013).

We expected that stimulus predictability and luminance contrast should interact in a multiplicative way for γ-synchronization. We reasoned that V1 does not have access to the image itself but needs to infer the image properties based on a potentially noisy representation of the image, encoded by sparse and variable lateral geniculate nucleus (LGN) inputs (Rao and Ballard, 1999). Hence, the inputs to V1 neurons should become more predictable/redundant for high luminance contrast and input drive (Peter et al., 2019). To investigate whether luminance contrast and predictability interacted in a multiplicative way (Figures 4C–4E), we first fit a multiple linear regression model. The regression predictors were structural predictability, luminance contrast and their interaction. This full regression model explained more variance in γ-synchronization than luminance contrast and predictability alone (Figure 4C). There was a significant interaction between luminance contrast and structural predictability (mean t-statistic contrast -1.42 ± 0.89, $p > 0.05$, t test; predictability -0.84 ± 0.78, $p > 0.05$; predictability × contrast 4.2 ± 1.2, $p < 0.001$). The interaction term of luminance contrast and structural predictability explained almost as much variance as the full regression model (Figure 4C), consistent with a multiplicative effect (see also Figure 4D).

To further investigate whether luminance contrast reflects the presence of a high amount of redundant information in the visual inputs to V1, we added Poisson noise to the original images, mimicking neural noise. We then quantified the resulting predictability as described above for each image (Figure 4F). This yielded a structural predictability value for the noisy images, indicating the extent to which a noisy RF input could be predicted by the noisy image context (predictability under noise) (Figure 4F). PUN correlated strongly with the luminance contrast in the original (i.e., noise-free) image (Figure 4F). Furthermore, PUN correlated more strongly with γ synchronization than the structural predictability computed over the original image (Figure 4G; similar results were obtained for Gaussian noise). By contrast, correlations of PUN with firing rates were similar to those with the original luminance contrast (Figure 4G). These findings support the idea that the inputs to V1 neurons become more predictable/redundant for high luminance contrast.

Finally, we examined other stimulus factors like spatial frequency and stimulus orientation. These factors explained little variance in γ, even though they had a strong effect on explained variance for firing rates (Figures 4C, **S4C**, and **S4G**). In addition, image focus, which was strongly correlated with luminance contrast, showed weaker correlations with γ than contrast and predictability (Figure S4H).

Synchronization is poorly accounted for by a feedforward network for object recognition

We wished to compare these analyzes to a standard approach in which neural activity is explained by the activations of units in a convolutional FF network for OR-CNN. Our OR-CNN instantiation was the Visual Geometry Group-16 (VGG-16) network. In contrast to the predictive neural network, which was trained using self-supervision to extract structural predictability, VGG-16 categorizes object images and is trained in a supervised way using labeled objects. V1 firing rates were relatively well accounted for by the OR-CNN, indicating a similarity between stimulus responses of V1 units and artificial units in middle layers of the OR-CNN (Figures 5A–5C), in agreement with previous work, Cadena et al. (2019). By contrast, the strength of γ-synchronization was relatively poorly explained by such a network, i.e., its stimulus selectivity was not well explained by the stimulus selectivity of OR-CNN units (Figure 5B). These results indicate that γ itself was poorly accounted for by a linear combination of low-level pixel values. By contrast, we have shown above that the strength of γ-synchronization was well explained by the structural predictability of those low-level pixel values (see Figures 2B and 2C), indicating structural predictability entails a non-linear computation.

To test this, we trained a *de novo* neural network to predict γ-synchronization from the image properties (Figure S5C). This yielded much stronger correlations ($r \approx 0.7$) that were close in magnitude to the correlations obtained by combining predictability and luminance contrast (Figures S5D and S5E). This means that predictability and luminance contrast are quite close to the performance of the black-box approach, which provides an estimate of the ceiling of possible performance.

Perceptual similarity and low versus high-level predictability

In Figures 2B–2D, we showed that late firing rates were not significantly correlated with structural predictability. Because firing rates were generally reduced for full as compared with small natural images (see Figure 7; Vinje and Gallant 2000), we wondered which factors determine the contextual modulation of firing rates (i.e., surround suppression). Structural predictability is based on measuring similarity between the predicted and the presented image using pixel-by-pixel correlations. This similarity function is closely related to the perceptual similarity measure SSIM, which is mathematically tractable and has an intuitive relationship to image compressibility (Figure 3B). Structural predictability does not distinguish between higher- and lower-level features. Yet, it is possible that neuronal signals may distinguish between predictability of low- and high-level features.

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

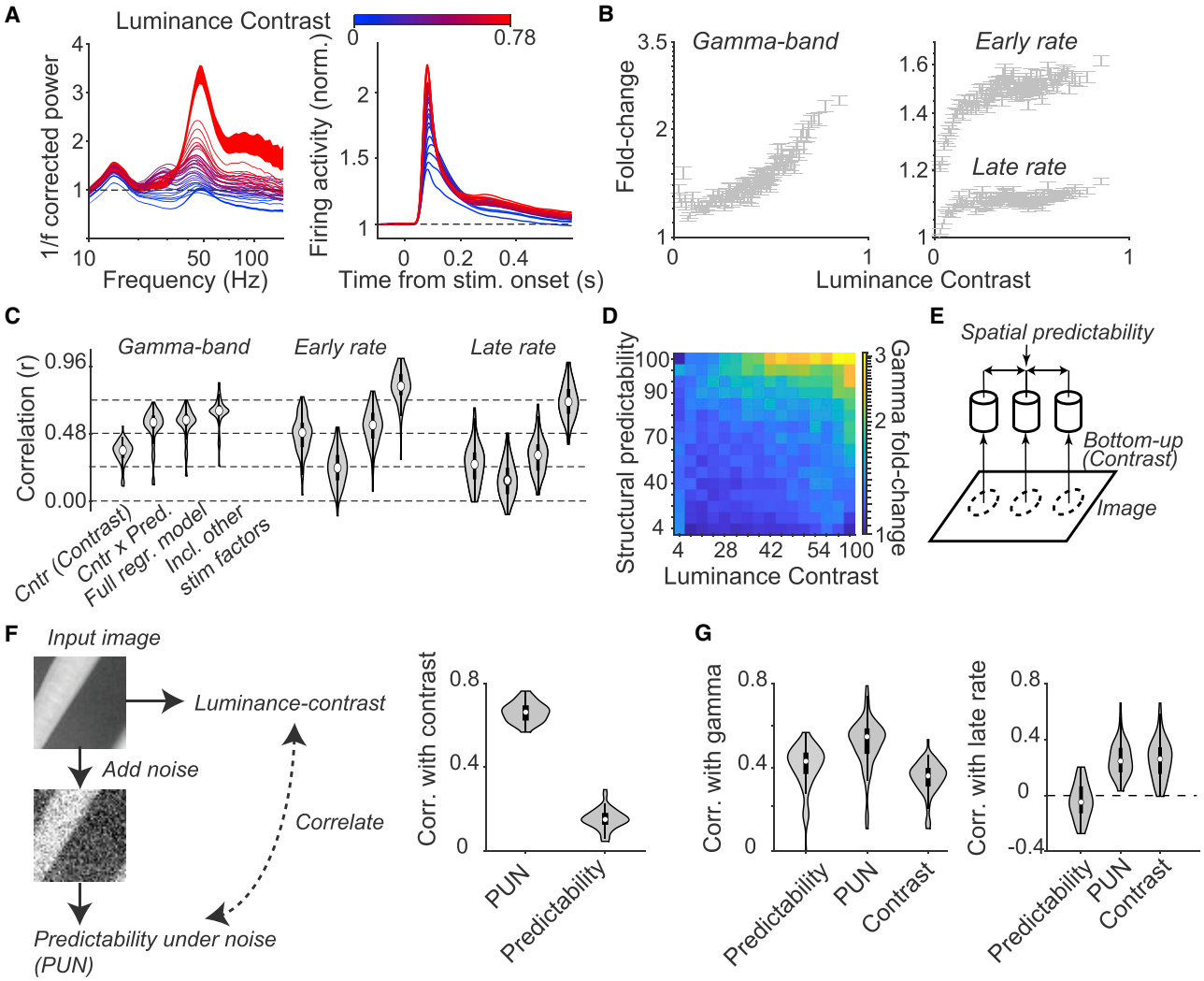

Figure 4. Dependence of neural activity on luminance contrast and predictability

(A) (Left) Average 1/*f*-corrected LFP power spectra (±1 SEM) for the highest and lowest level of luminance contrast (root means square contrast , see STAR Methods), for monkey H. (Right) As left, but for multi-unit firing rates.

(B) (Left) Average γ-peak amplitude versus luminance contrast. (Right) Same for early (50–150 ms) and late MU firing rates (200–600 ms).

(C) Average correlation across sites of γ and firing rate with image factors. Left-to-right: (Ci) Luminance contrast (Cntr); (Cii) the product of contrast and stimulus predictability (Cntr × Pred.); (Ciii) Pred, Cntr, Pred. × Cntr interaction (Full regression (regr.) model); and (Civ) a model with additional low-level features (Including (Incl.) other stim factors), namely spatial frequency, luminance and orientation (see STAR Methods). Correlations were computed for each recording site separately, across all images presented across sessions. All correlations were significantly different from zero ($p<0.001$, paired t test). For γ, the difference between (Cntr × Pred.) and (Full regr. model) was significant ($p<0.001$), but the difference between (Full regr. model) and (Incl. other stim factors) was not ($p>0.05$). For early rates, all comparisons were significant ($p<0.05$). For late rates, all comparisons except for (Full regr. model) versus (Cntr × Pred.) were significant at $p<0.05$.

(D) γ fold-changes for different levels of luminance contrast and structural predictability.

(E) Illustration of interaction between predictability and bottom-up inputs.

(F) Derivation of the PUN measure (predictability under noise, left). We added Poisson noise to each luminance value and then computed the structural predictability for the noise-corrupted image, yielding the PUN measure. PUN was strongly correlated to the original luminance contrast in the center RF (right).

(G) (Left) PUN correlated more strongly with γ synchronization than predictability and luminance contrast ($p<0.001$ for both, paired t test). (Right) For late firing rates, correlations with PUN were weaker than for luminance contrast ($p<0.001$). For a comparison of baseline-corrected γ-power with 1/*f*-corrected γ-power, an analysis of reliability of the predictors, and analysis of other stimulus factors, see Figure S4. Data in (A and B) are represented as mean ± SEM. See also Figure S4.

Because perceptual similarity measures that are based on pixel-by-pixel correlations like SSIM do not distinguish between higher and lower-level image features, we turned to a newer method for perceptual similarity, learned perceptual image patch similarity (LPIPS). LPIPS compares two images based on activations in each layer of an OR-CNN (Figure 6A; Zhang et al., 2018). The different OR-CNN layers represent different levels of image feature abstraction. LPIPS showed a stronger correlation with human perceptual similarity judgments than SSIM (Figure 6B), consistent with Zhang et al. (2018).

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

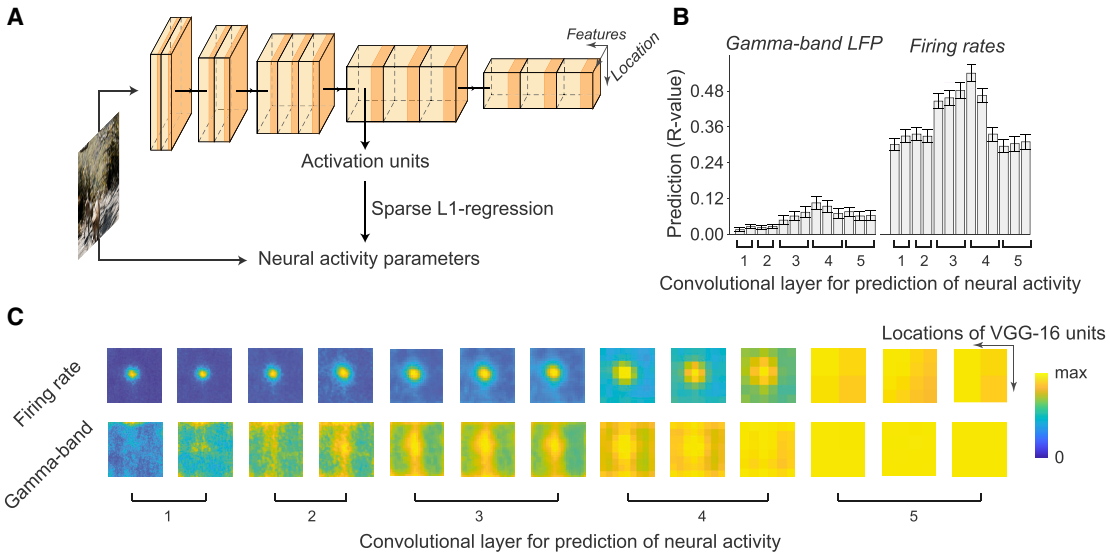

Figure 5. A feedforward neural network for object recognition explains firing rates relatively well, but poorly accounts for γ-synchronization

(A) For each recording site, we determined different neural activity parameters. The image patch centered on the RF of the recording site was then passed into the CNN for object recognition (OR-CNN; in this case the VGG-16), and we computed the activation of every OR-CNN artificial neuron (AN) whose RF overlapped with the recording site. Sparse L1-regression with cross-validation was used to predict neural activity from OR-CNN ANs with RFs at the center of the image.

(B) Regression prediction accuracy of different neural activity parameters depending on OR-CNN layer. Data are represented as mean ± SEM. Regression prediction accuracy for late (200–600 ms) firing rates was significantly higher for middle (5–9) than early (1–4) and deep (10–13) convolutional layers ($p< 0.001$, paired t test). For γ, regression prediction accuracy was significantly higher for middle ($p<0.001$) and deep ($p<0.05$) than early layers. For early rates and β see Figures S5A and S5B.

(C) Prediction accuracy depending on the RF location of OR-CNN ANs in the image. In this case, we predicted neural activity from all units in a 3 × 3 image using sparse L1-regression. Shown are the prediction weights, which reveal circular RFs for firing rates already in the earliest layer. See also Figure S5.

We then used the global LPIPS measure, which summarizes over OR-CNN layers, to compare the ground-truth and predicted image (using the algorithm detailed in Figure 1), yielding "LPIPS-predictability." LPIPS-predictability captures not only how well low-level, but also how well higher-level features of the stimulus in the RF are predicted by context. We found that V1 rates in the late response phases were negatively and monotonically related to LPIPS-predictability (Figure 6C), in contrast to the absence of a significant correlation with structural predictability (Figure 2).

These observations suggest the importance of deeper OR-CNN layers in explaining neural activity and human perceptual similarity judgments. To quantify CNN-layer-specific measures of image similarity, we measured the similarity between OR-CNN activation patterns resulting from two image inputs using the Euclidean distance between activation patterns, a measure of representational differences between image content (content similarity). In addition, we computed OR-CNN-based structural similarity as the Pearson correlation between OR-CNN activation patterns, in analogy to pixel-wise structural similarity, thus weighting spatial correlations more than image content. OR-CNN-based content and structural similarity were most strongly correlated to human perceptual similarity judgments in deep layers (Figure 6B). We then distinguished between low and high-level predictability by computing layer-specific OR-CNN-based content and structural predictability. We found that the predictability of higher-level OR-CNN features showed a stronger negative correlation with image salience than the predictability of lower-level OR-CNN features (Figure S6D). By contrast, image compressibility correlated best with the activity in early, rather than deep layers of the OR-CNN (Figure S6A).

For both OR-CNN-based stimulus predictability measures, we found that rates were negatively related to predictability, and that the magnitude of the correlation between stimulus predictability and V1 firing rates became stronger toward the deeper layers of the OR-CNN (Figure 6D; showing the magnitude of the average correlation). Thus, high-level predictability was the most reliable variable explaining V1 firing rates. We observed the opposite pattern for V1 γ-synchronization (for β see Figure S6Ei): V1 γ was strongly and positively correlated with OR-CNN-based structural and content predictability in the early layers of the OR-CNN (Figure 6D) but weakly correlated with content and OR-CNN-based structural predictability in the deeper layers of the OR-CNN. Accordingly, the correlation with OR-CNN-based predictability decreased across layers (Figure 6D). Including OR-CNN-based structural and content predictability from deeper layers did not explain further variance in V1 γ-synchronization, whereas it did for V1 firing rates (Figure S6B).

These findings indicate that γ-synchronization and firing intensity reflect distinct aspects of stimulus predictability, and with opposite correlations of opposite signs: firing rates decrease with the predictability of high-level features, whereas γ-synchronization increases with low-level predictability.

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

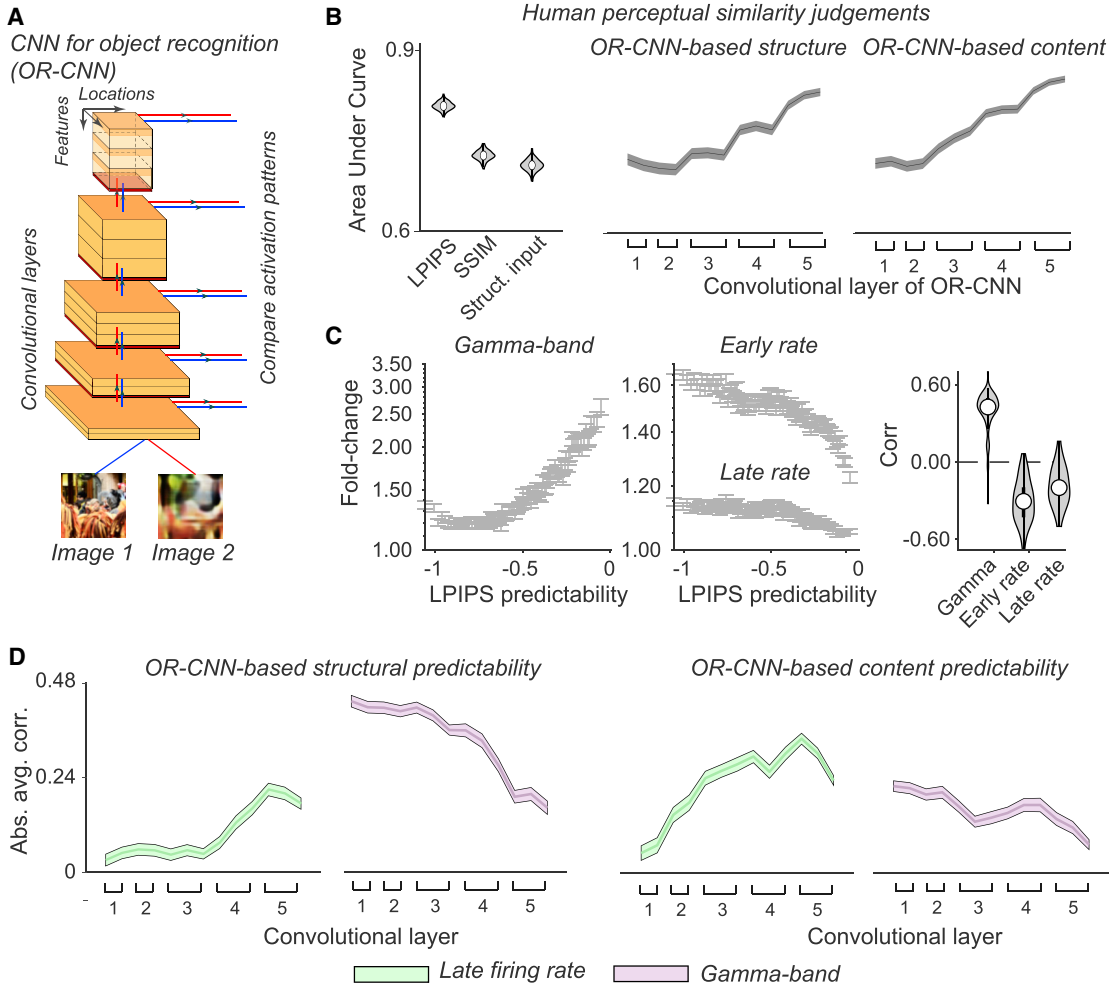

Figure 6. Firing rates reflect high-level stimulus predictability, gamma reflects low-level stimulus predictability

(A) OR-CNN network (VGG-16) used to define low- and high-level stimulus similarity and predictability. Responses of artificial units (ANs) in different layers OR-CNN layers were computed for two images at a time. For each layer, we computed two similarity measures: (A1) content similarity, which is based on Euclidean distance; (A2) OR-CNN-based structural similarity, which was computed as the Pearson correlation across locations for each AN separately and then averaging these correlations across ANs.

(B) (Left) Average AUC value for (B1) LPIPS, (B2) SSIM, (B3) structural correlations, as used for Figures 1 and 2. Learned perceptual image patch similarity (LPIPS) is a perceptual similarity measure based on OR-CNNs (Zhang et al., 2018). LPIPS had higher AUC values than the other measures ($p<0.001$, paired t test). (Right) Structural and content similarity versus human perceptual similarity. AUC increased significantly with layer depth for both structure and content ($r = 0.93$ and $r = 0.98$, $p<0.001$ for both).

(C) (Left) Firing rates and neural synchronization versus LPIPS-predictability. The input (ground-truth) and predicted image-patch were compared using the OR-CNN network, yielding LPIPS-predictability. (Right) Correlations across all images, averaged across recording sites ($p<0.001$) for all variables).

(D) Correlation of late firing rates and peak γ-power with OR-CNN-based content and structural predictability across OR-CNN layers. See Figure S6C for example images with different levels of low- and high-level content predictability. Note that, we first computed average correlations and show here the absolute average value of these correlations, but that correlations were positive for γ and negative for firing rates. Late firing rates showed a significant increase in absolute correlation across OR-CNN layers, both for OR-CNN-based structural and content predictability (structure: $r = 0.85$; content: $r = 0.81$, $p<0.001$ for both). By contrast, γ showed a significant decrease for both (structure: $r = 0.9$; content: $r = 0.84$, $p<0.001$). The average correlation (across layers) for OR-CNN-based structural predictability was significantly higher than OR-CNN-based content predictability for γ ($p<0.001$, paired t test). For firing rates, the average correlation with OR-CNN-based content predictability was higher than for OR-CNN-based structural predictability ($p<0.001$, paired t test). Data in (B–D) are represented as mean ± SEM. See also Figure S6.

Experimental dissociation of firing rates and γ-synchronization

To further dissociate firing rates from γ-synchronization we designed an additional experiment, in which stochastic textures (brown, white, and pink noise) were presented in two conditions: center-surround match (same noise in RF and surround) and mismatch (different textures in RF and surround) (Figure 7A). This paradigm can be compared with the classic figure-ground paradigm (Lamme and Spekreijse, 1998), but in our case, the figure region has a small size (the RF region). Because the texture

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

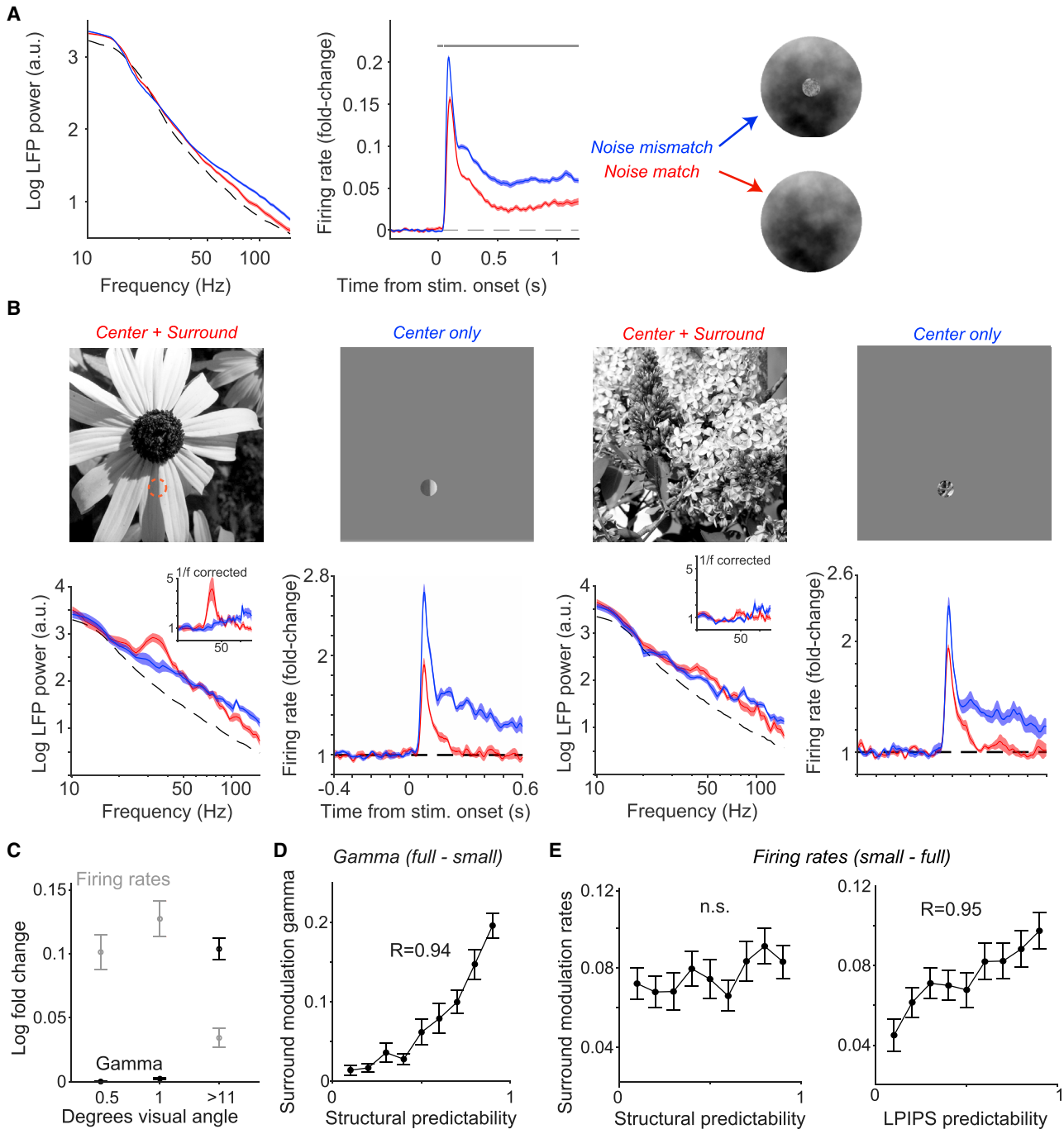

Figure 7. Firing rates and gamma show distinct modulations by spatial context

(A) Center-surround mismatch paradigm with noise stimuli and center-only versus full stimuli. Stimuli had white, pink, or brown noise in the 1 dva center, and white, pink, or brown noise in the surround. Stimuli (6 dva) were centered on the recording site's RF. Only recording sites within 0.25 dva of the stimuli center were analyzed. (Left) LFP power spectra. Note that the broadband increase in LFP power at high frequencies is typical for spike bleed-in Ray and Maunsell (2011). (Dashed line) Baseline pre-stimulus period. (Right) Normalized MU firing rates. Firing rates were higher for noise-mismatch stimuli (gray bar: $p<0.05$, t test, n = 24 recording sites).

(B) (Left and right): Examples of surround suppression for image that show either clear gamma synchronization (left) or no clear peak in the gamma-range.

(C) Comparison of gamma-amplitude and late firing rates for different stimulus sizes (log_{10}-fold-change for both). Suppression for early firing intensity was significantly weaker than for late firing intensity (paired t test, $p<0.001$).

(D) Increase in γ-synchronization for full compared with small images as a function of structural predictability.

(E) Surround modulation in firing rates (small minus full) as a function of structural predictability and LPIPS-predictability. Data in (A–C) are represented as mean ± SEM. See also Figure S7.

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

stimuli had low structural predictability, we did not expect to find any γ-synchronization. Indeed, LFP spectra did not show γ-peaks for any of the conditions (Figure 7A). However, the texture stimuli differed in terms of higher-level content predictability since artificial neurons (ANs) in deeper layers of OR-CNNs have texture selectivity. Accordingly, firing rates were substantially higher for texture-mismatch stimuli than homogeneous textures (Figure 7A). These findings further support the conclusion that firing rates and synchronization properties are modulated by distinct aspects of spatial predictability.

Stimulus size dependence and surround modulation

Finally, we investigated whether the surround modulation of neural activity, i.e., the difference between responses to small and large stimuli, was explained by stimulus predictability. We did this for two reasons: first, it allowed us to dissociate the dependence on context from local stimulus properties, keeping the local properties constant while exclusively manipulating the surround. Second, we wanted to investigate if surround modulation itself was determined by stimulus predictability.

We recorded additional sessions in which images were presented either in full ("center+surround") or within a small (0.5–1 dva) aperture ("center-only") centered on the RFs of the recorded neurons. Firing rates were consistently reduced in the center+surround condition (Figures 7B, 7C, and S7A), consistent with previous work (Vinje and Gallant, 2000; Coen-Cagli et al., 2015). Surround suppression was stronger for late than for early firing rates (pairwise t test, $p<0.001$, early: 0.043 ± 0.009, late: 0.08 ± 0.008) consistent with the dependence of surround suppression on horizontal and top-down feedback (FB) (Angelucci et al., 2017). In contrast to firing rates, γ-synchronization was strongly reduced in the center-only condition (Figures 7C and S7A). Thus, the emergence of γ required that there was a predictable RF "surround," i.e., that the natural image extended beyond 1 dva with a predictable context, rather than a gray screen that did not match the center stimulus (Figure 7C), consistent with previous findings using artificial stimuli (Peter et al., 2019; Gieselmann and Thiele, 2008).

Surround suppression of firing rates was sometimes associated with strong γ-synchronization and in other cases occurred in the absence of a clear γ peak (Figure 7B). Surround modulation of γ (i.e., an increase for large stimuli) became increasingly stronger with structural predictability (Figure 7D). By contrast, the surround modulation of firing rates (i.e., the decrease for large stimuli) did not show a significant trend with structural predictability (Figure 7E, left). The surround modulation of firing rates increased monotonically with LPIPS-predictability, however (Figure 7E, right). We further compared LPIPS-predictability with previously developed models for firing rate surround modulation based on linear correlations between Gabor-like simple cells (Coen-Cagli et al., 2012, 2015). We show that LPIPS-predictability explained surround modulation better than this state-of-the-art model (Figures S7B and S7C). Thus, the stimulus predictability measures developed here allow for a continuous regression of the surround modulation of firing rates and γ synchronization, which are modulated by distinct types of stimulus predictability.

DISCUSSION

Summary

We find that V1 firing rates and synchronization have distinct relationships to the spatial predictability of visual stimuli, suggesting complementary roles in predictive processing.

Predictability for natural scenes was quantified using a self-supervised neural network. This network was trained on a large number of images, and generated predictions of likely visual stimuli falling into the neuronal RFs. We reasoned that the predictive neural network has learned the statistical structure of the stimulus set and developed a similar internal model as the primate visual system to generate predictions. We then defined two kinds of predictability measures: first, by comparing the predicted RF input to the actual RF input, we defined structural predictability as the extent to which the precise structure of a stimulus in the V1 RF can be predicted by the context. Second, we distinguished between the predictability of lower- and higher-level features using a OR-CNN. OR-CNNs are standard models of neural responses in the primate ventral stream and have recently been used to compute perceptual image similarity by comparing images in terms of their activation across OR-CNN layers (Zhang et al., 2018). Due to the influence of deeper OR-CNN layers, CNN-based image similarity provides a closer match to human perceptual similarity judgments than structural similarity, Zhang et al. (2018). We defined lower- and higher-level predictability by comparing actual and predicted RF input in terms of OR-CNN features. Higher-level predictability showed a stronger negative correlation with image salience than lower-level predictability, demonstrating the behavioral relevance of higher-level predictability. Our main findings are the following:

(1) Structural predictability in natural scenes is a key determinant of γ-synchronization (30–80 Hz). Consequently, γ-synchronization emerges for predictable stimuli that have low dimensionality, can be efficiently encoded (compressibility), and have low image salience. Surprisingly, firing rates are only weakly modulated by structural predictability and image compressibility. An unexpected observation was that stimuli with weak structural predictability, esp. with nature content, induce a late-onset (>500–600 ms) β-rhythm in the V1 LFP (see further below).

(2) The main factor determining a decrease in V1 firing rates is the contextual predictability of "higher-level" features of stimuli falling into the RF. By contrast, the key determinant of increases in γ-synchronization is the contextual predictability of "lower-level" OR-CNN features. Consequently, contextual predictability of stochastic textures decreases firing rates, but does not modulate γ-synchronization. In sum, not only does the sign of the relationship to spatial predictability differ between firing rates and γ-synchronization (negative and positive correlations, respectively), but also the level of spatial predictability they are modulated by (higher and lower, respectively).

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

Distinct effects of predictability on firing rates and synchronization

Neuronal populations may be modulated by predictability in two distinct ways: (1) through changes in firing rates and (2) through distinct patterns of correlated firing that result from the recurrent interactions within and between neuronal populations (Bastos et al., 2020; Vinck and Bosman, 2016; Singer, 2021). In this study, we focused on a particular form of correlated firing, namely local rhythmic synchronization in the γ-frequency range. This kind of synchronization is, spectrally speaking, a narrow-band phenomenon and is therefore visible in the LFP (Buzsáki, 2006; Pesaran et al., 2018). γ-synchronization is thought to reflect local interactions between inhibitory and excitatory neurons. In addition, cat and primate V1 contains a unique class of excitatory pacemaker neurons that may explain the prominence of the V1 γ-rhythm compared with other cortical areas (Onorato et al., 2020; Gray and McCormick, 1996).

Overall, our findings agree with the hypothesis that V1 γ-synchronization reflects spatial predictability (Vinck and Bosman, 2016). We further show that this relationship pertains specifically to the predictability of low-level features (structural predictability). However, our findings appear incompatible with earlier studies that found γ-synchronization to be positively related to prediction errors (Arnal et al., 2011; Bastos et al., 2020; Chao et al., 2018; Bauer et al., 2014). Notably, these studies used temporal predictability, whereas the present study used spatial predictability, and the generalization of our findings to temporal predictability remains to be demonstrated (see Vinck and Bosman, 2016; Peter et al., 2019; Canales-Johnson et al., 2021 for further discussion). We further note that there are different flavors of predictive coding theories (Rao and Ballard, 1999; Friston, 2008; Mumford, 1992; Keller and Mrsic-Flogel, 2018; de Lange et al., 2018; Heeger, 2017; Singer, 2021). The dependence of V1 γ-synchronization on the structure of artificial stimuli (Figure S3) as well as the strong γ responses to homogeneous colored surfaces (Peter et al., 2019; Shirhatti and Ray, 2018; Rols et al., 2001) further support our conclusion that γ-synchronization reflects structural predictability and emerges for simpler forms of stimulus continuities across space, e.g., edges, line elements, and surface color (Peter et al., 2019).

To our surprise, firing rates correlated very weakly with structural predictability and image compressibility but showed a relatively strong decrease with the predictability of higher-level OR-CNN features. Previous work has shown that the activity in middle layers of OR-CNNs provides a reasonable model for V1 activity; however, this model leaves substantial variance unexplained (Cadena et al., 2019) (see also Figure 5). Thus, our findings provide a quantitative approach to predict V1 firing rates from natural images based on predictability, which provides complementary explanatory power compared with previous approaches based on OR-CNNs. These findings on predictability and OR-CNNs suggest that V1 firing rates result from two kinds of computations: (1) they encode features that are relevant for core visual tasks such as object recognition and (2) they increase when those features are not predicted by or stand out from the context. Importantly, the goal of OR-CNNs is to encode and extract categorical information in an invariant manner, not to predict or reconstruct the sensory inputs. Thus, it makes sense that firing rates were weakly modulated by the extent to which the structure of the image can be predicted by a low-dimensional representation.

Importantly, if stimulus manipulations increase both low- and high-level predictability, then they can lead to opposite changes in γ-synchronization (i.e., an increase) and firing rates (i.e., a decrease) (Peter et al., 2019). An example of this is increasing the size of a stimulus by adding a predictive surround, which typically leads to increases in γ-synchronization and decreases in firing rates for neuronal populations at the stimulus center (Peter et al., 2019; Gieselmann and Thiele, 2008). Another example is a center-surround mismatch in stimulus orientation, which enhances firing rates and suppresses γ-synchronization for neuronal populations at the stimulus center (Bair et al., 2003; Veit et al., 2017). Importantly, compared with cross-oriented center-surround gratings, iso-oriented gratings have stronger predictability both in terms of lower-level features (i.e., a continuity of line elements) and higher-level features (i.e., similar stimulus orientation and spatial frequency).

In sum, γ-synchronization increases with low-level predictability, whereas firing rates decrease with high-level predictability. A possible explanation for this dissociation may be the following: γ-synchronization likely depends on local interactions among excitatory/inhibitory neurons via horizontal, patchy connections, which have a limited spatial reach of a few millimeters (Vinck and Bosman, 2016; Veit et al., 2017; Lowet et al., 2017; Rockland and Lund, 1983; Gilbert and Wiesel, 1983; Gray et al., 1989) (note that V2 FB may in addition modulate γ-synchronization [Hartmann et al., 2019; Vinck and Bosman, 2016]). As a result, γ-synchronization may decrease mainly due to local discontinuities in stimuli. By contrast, firing rates can increase due to mismatches across larger regions of space, mediated by both long- and short-range top-down FB (Angelucci et al., 2017; Keller et al., 2020; Kirchberger et al., 2021; Keller and Mrsic-Flogel, 2018). This longer-range FB can mediate the computation of mismatches in higher-level features across a larger region of space.

Finally, although V1 spiking rhythmicity was generally very weak for non-predicted stimuli, we observed a clear β-rhythm in the LFP for these stimuli. This β-rhythm emerged almost exclusively for pictures of nature content and did not correlate with luminance contrast. Previous work suggests that β reflects cortical top-down FB from parietal and frontal areas (Bressler et al., 2007; Buschman and Miller, 2007; Bastos et al., 2015; Gregoriou et al., 2009). This is consistent with the late onset of β after 500–600 ms, given that late, sustained activity is very weak in area V1 compared with other areas in parietal and frontal cortex. Overall, the distinct properties of β-synchronization, as well as the lack of a gradual frequency shift from γ to β frequencies, suggest that the β-rhythm is a phenomenon that is distinct from the V1 γ-rhythm. More data are needed on the generative source of this β-rhythm and its relationship to top-down FB and predictions (Bastos et al., 2020, 2012) before a functional interpretation can be given.

Functional interpretations

Next, we will discuss potential roles of γ-synchronization and firing rate modulations in several domains.

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

EFFICIENT CODING

Our findings strongly suggest that V1 γ-synchronization is associated with efficient encoding of the image. A previous computational model has studied how a population of neurons can encode a 1D stimulus (Chalk et al., 2016). Optimal encoding of a 1D stimulus requires (1) intermediate levels of γ-synchronization (Chalk et al., 2016), (2) sparsely firing cells that spike in a small fraction of γ-cycles, and (3) stochastic γ-oscillations, as observed in area V1 (Spyropoulos et al., 2020; Burns et al., 2011). Thus, V1 γ-synchronization may represent a dynamic regime that allows for efficient encoding (i.e., data compression) of low-dimensional, highly redundant images. In contrast, when V1 neurons represent non-redundant information and receive heterogeneous inputs, efficient encoding entails that correlations between neurons need to be avoided and γ-synchronization should decrease. A similar explanation may account for the fact that in mouse LGN, γ-synchronization occurs for diffuse light stimulation rather than for structured stimuli (Saleem et al., 2017; Schneider et al., 2021).

PLASTICITY

It is known that synchronization can have important consequences for the induction of synaptic plasticity (Galuske et al., 2019; Sejnowski and Paulsen, 2006; Stopfer et al., 1997), which is mediated by mechanisms such as spike-timing-dependent plasticity (Wespatat et al., 2004; Sjöström et al., 2001; Sejnowski and Paulsen, 2006; Vinck et al., 2010; Galuske et al., 2019; Anisimova et al., 2021). Bursts of γ oscillations could therefore maintain or strengthen synaptic connections between neurons that, on average, predict each other's visual inputs, and thereby contribute to self-supervised learning of spatiotemporal natural image statistics.

INTER-AREAL COMMUNICATION

We found that γ-synchronization is enhanced for predicted stimuli and decreases with image salience. A possible interpretation based on predictive coding models is that γ-synchronization reduces FF information propagation. In contrast to this interpretation, previous work has shown that FF Granger-causality from V1 to higher areas is strongly associated with γ-synchronization (Bosman et al., 2012; Bastos et al., 2012, 2015; van Kerkoerle et al., 2014). However, a recent study argues that a strong γ-source in V1 will naturally give rise to FF Granger-causality with LFP signals in higher areas, without necessarily having a direct functional consequence for information transmission (Schneider et al., 2021). Thus, the critical question is the precise effect of γ-synchronization on cells in downstream targets. Previous work suggests that fast-spiking interneurons respond more strongly to γ-frequency inputs than excitatory neurons, as excitatory neurons tend to have low-pass filtering properties (Pike et al., 2000; Buzsáki and Schomburg, 2015; Cardin et al., 2009; Hasenstaub et al., 2005). As a result, γ-synchronization may recruit strong FF inhibition in a receiving area (Schomburg et al., 2014), instead of activating and entraining excitatory neurons. Consistent with this interpretation, several studies have shown that γ-synchronization in a sending area (e.g., LGN and V1) induces phase-locking only in layer IV of the receiver (e.g., V1 and V2), without propagation to layers II/III (Zandvakili and Kohn, 2015; Schneider et al., 2021) (note that other studies have reported inter-areal spike phase-locking without selecting particular layers or neuron types [Engel et al., 1991; Grothe et al., 2012]).

These possible interpretations seem difficult to reconcile with the hypothesis that inter-areal phase-synchronization between γ-rhythms can selectively gate communication according to cognitive demands such as attention (Bosman et al., 2012; Rohenkohl et al., 2018; Grothe et al., 2012). Here, we observed substantial variability in V1 γ-synchronization across stimuli, which may result in substantial variability in inter-areal γ-coherence with areas V2 and V4 (Schneider et al., 2021; Roberts et al., 2013). Given such variability, it remains an open problem how attention can be generically mediated via selective inter-areal synchronization, and how stimuli with weak γ-synchronization are communicated to higher areas (for further discussion see Hermes et al. 2015; Ray and Maunsell 2015; Brunet et al. 2014; Akam and Kullmann 2012). Our data show that the variability in V1 γ-synchronization across natural images is not explained by the extent to which a stimulus is "visible" or salient; note that the negative correlation of γ with salience is explained by the strongly negative correlation of salience with predictability (in line with Li [2002]). Furthermore, the effect of attention on γ-synchronization in V1 varies across studies and does not show a systematic tendency in any direction (van Kerkoerle et al., 2014; Chalk et al., 2010; Das and Ray, 2018; Ferro et al., 2021; Buffalo et al., 2011). By contrast, V1 firing rates are typically enhanced both for stimuli that are attended (Chalk et al., 2010; Buffalo et al., 2010; van Kerkoerle et al., 2014) and that contain a center-surround mismatch in higher-level features.

MISMATCH SIGNALS

These firing rate mismatch signals can play numerous important functions. They can be important for learning, given the well-established role of error signals in learning (Bellec et al., 2020). Furthermore, mismatch signals can attract attention to stimuli that are salient and stand out from the background, guiding eye movements (Li, 2002). Mismatch signals may also enrich the heterogeneity of features that are computed within V1. In this context, both the local RF properties of a V1 neuron and its contextual modulation can be seen as a "difference" operator: the local classical RF properties result from a local difference operation (e.g., a Gabor-like filter), whereas contextual modulation reflects a difference operation on a larger spatial scale. Mismatch signals on a larger spatial scale may benefit performance on tasks such as object recognition and contribute to scene segmentation and perceptual grouping. Indeed, it is known that in mice, performance on scene segmentation tasks depends on the contextual modulation of V1 firing rates (Kirchberger et al., 2021). By contrast, the relationship of γ-synchronization with perceptual grouping has been debated (Roelfsema et al., 2004; Palanca and DeAngelis, 2005; Shadlen and Movshon, 1999; Singer, 1999; Lima et al., 2010). Our data suggest that rates and synchronization may carry

complementary information about segmentation. Whereas firing rates are increased by spatial discontinuities in higher-level features (defined by OR-CNNs), such as stochastic textures, γ-synchronization decreases with local discontinuities in lower-level structure (e.g., edges, color). Finally, recent work has suggested that inserting a V1-like layer into OR-CNNs may improve object recognition (Marques et al., 2021; Dapello et al., 2020). Here, we presented a quantitative method to predict contextual modulation of V1 firing rates. It needs to be investigated whether including a similar contextual modulation of firing rates in OR-CNNs benefits object recognition. Likewise, including γ-oscillatory mechanisms in recurrent neural networks with similar properties as observed here may benefit task performance and efficient coding.

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
- METHOD DETAILS
  - Surgical procedures
  - Behavioral task
  - Recordings
  - Visual stimulation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data analysis
  - Identification of object boundaries and man-made vs. nature pictures
  - Deep Neural Networks methods
  - Preprocessing of training stimuli
  - Architecture of deep neural network for inpainting
  - Loss function for training
  - Training and Hyperparameter Optimization
  - Image Statistics
  - Visual Saliency
  - Human perceptual similarity
  - Predicting neural activity from VGG-16 activations
  - Prediction from VGG-16 activations: gamma-net

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2022.01.002.

ACKNOWLEDGMENTS

We thank Michael Schmid and Richard Saunders for implantation of monkey H, and C. Bosman, S. Neuenschwander, Q. Perrenoud, C. Richter, H. Kennedy, and W. Maass for FB. This project was supported by an ERC Starting Grant (SPATEMP, EU), a BMBF (Germany) Grant to M.V. (Computational Life Sciences, project BINDA, 031L0167), Nvidia (two GPUs), an HFSP Research Grant, (RGP0044/2018-Orban to W.S., France) and a DFG Reinhart Koselleck-Projekt (GZ: SI 505/22-1 to W.S., Germany).

AUTHOR CONTRIBUTIONS

Conceptualization and design: C.U., A.P., and M.V. Data analysis: C.U. and M.V. Design of neural networks and related methods: C.U. and M.V. Implementation of neural networks: C.U. Recordings: A.P., C.U., R.R., and J.K.-L. Software: C.U., A.P., A.L., and K.A.S. Surgery: J.K.-L., W.B., and W.S. Materials and reagents: W.S., P.F., and M.V. Supervision: M.V. Writing – original draft: C.U., A.P., and M.V.

DECLARATION OF INTERESTS

P.F. has a patent on thin-film electrodes and is beneficiary of a respective license contract with Blackrock Microsystems (Salt Lake City, UT, USA). P.F. is a member of the Scientific Technical Advisory Board of CorTec (Freiburg, Germany) and is managing director of Brain Science (Frankfurt am Main, Germany).

Received: May 14, 2021
Revised: November 22, 2021
Accepted: January 4, 2022
Published: February 3, 2022

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al. (2015). TensorFlow: large-scale machine learning on heterogeneous systems (TensorFlow). https://www.tensorflow.org/.

Akam, T.E., and Kullmann, D.M. (2012). Efficient "communication through coherence" requires oscillations structured to minimize interference between signals. PLoS Comp. Biol. *8*, e1002760.

Angelucci, A., Bijanzadeh, M., Nurminen, L., Federer, F., Merlin, S., and Bressloff, P.C. (2017). Circuits and mechanisms for surround modulation in visual cortex. Annu. Rev. Neurosci. *40*, 425–451.

Anisimova, M., van Bommel, B., Mikhaylova, M., Wiegert, J.S., Oertner, T.G., and Gee, C.E. (2021). Spike-timing-dependent plasticity rewards synchrony rather than causality. bioRxiv. bioRxiv. https://doi.org/10.1101/86336.

Araujo, A., Norris, W., and Sim, J. (2019). Computing receptive fields of convolutional neural networks (Distill). https://distill.pub/2019/computing-receptive-fields/.

Arık, S.O., Jun, H., and Diamos, G. (2018). Fast spectrogram inversion using multi-head convolutional neural networks. IEEE Signal Process. Lett. *26*, 94–98.

Arnal, L.H., Wyart, V., and Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat. Neurosci. *14*, 797–801.

Bair, W., Cavanaugh, J.R., and Movshon, J.A. (2003). Time course and time-distance relationships for surround suppression in macaque V1 neurons. J. Neurosci. *23*, 7690–7701.

Bastos, A.M., Lundqvist, M., Waite, A.S., Kopell, N., and Miller, E.K. (2020). Layer and rhythm specificity for predictive routing. Proc. Natl. Acad. Sci. USA *117*, 31459–31469.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., and Friston, K.J. (2012). Canonical microcircuits for predictive coding. Neuron *76*, 695–711.

Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J.R., De Weerd, P., Kennedy, H., and Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron *85*, 390–401.

Bauer, M., Stenner, M.-P., Friston, K.J., and Dolan, R.J. (2014). Attentional modulation of alpha/beta and gamma oscillations reflect functionally distinct processes. J. Neurosci. *34*, 16117–16125.

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., and Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. Nat. Commun. *11*, 3625.

Bengio, Y., Courville, A.C., and Vincent, P. (2012). Unsupervised feature learning and deep learning: a review and new perspectives, arXiv, arXiv:1206.5538v1.

Biederlack, J., Castelo-Branco, M., Neuenschwander, S., Wheeler, D.W., Singer, W., and Nikolić, D. (2006). Brightness induction: rate enhancement and neuronal synchronization as complementary codes. Neuron *52*, 1073–1083.

Bosman, C.A., Schoffelen, J.M., Brunet, N., Oostenveld, R., Bastos, A.M., Womelsdorf, T., Rubehn, B., Stieglitz, T., De Weerd, P., and Fries, P. (2012). Attentional stimulus selection through selective synchronization between monkey visual areas. Neuron *75*, 875–888.

Bressler, S.L., Richter, C.G., Chen, Y., and Ding, M. (2007). Cortical functional network organization from autoregressive modeling of local field potential oscillations. Stat. Med. *26*, 3875–3885.

Brunet, N., Vinck, M., Bosman, C.A., Singer, W., and Fries, P. (2014). Gamma or no gamma, that is the question. Trends Cogn. Sci. *18*, 507–509.

Buffalo, E.A., Fries, P., Landman, R., Buschman, T.J., and Desimone, R. (2011). Laminar differences in gamma and alpha coherence in the ventral stream. Proc. Natl. Acad. Sci. USA *108*, 11262–11267.

Buffalo, E.A., Fries, P., Landman, R., Liang, H., and Desimone, R. (2010). A backward progression of attentional effects in the ventral stream. Proc. Natl. Acad. Sci. USA *107*, 361–365.

Burns, S.P., Xing, D., and Shapley, R.M. (2011). Is gamma-band activity in the local field potential of V1 cortex a "clock" or filtered noise? J. Neurosci. *31*, 9658–9664.

Buschman, T.J., and Miller, E.K. (Mar. 2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science *315*, 1860–1862.

Buzsáki, G. (2006). Rhythms of the Brain (Oxford University Press).

Buzsáki, G., and Draguhn, A. (2004). Neuronal oscillations in cortical networks. Science *304*, 1926–1929.

Buzsáki, G., and Schomburg, E.W. (2015). What does gamma coherence tell us about inter-regional neural communication? Nat. Neurosci. *18*, 484–489.

Cadena, S.A., Denfield, G.H., Walker, E.Y., Gatys, L.A., Tolias, A.S., Bethge, M., and Ecker, A.S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Comp. Biol. *15*, e1006897.

Canales-Johnson, A., Teixeira Borges, A.F.T., Komatsu, M., Fujii, N., Fahrenfort, J.J., Miller, K.J., and Noreika, V. (2021). Broadband dynamics rather than frequency-specific rhythms underlie prediction error in the primate auditory cortex. J. Neurosci. *41*, 9374–9391.

Cardin, J.A., Carlén, M., Meletis, K., Knoblich, U., Zhang, F., Deisseroth, K., Tsai, L.-H., and Moore, C.I. (2009). Driving fast-spiking cells induces gamma rhythm and controls sensory responses. Nature *459*, 663–667.

Cavanaugh, J.R., Bair, W., and Movshon, J.A. (2002). Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. J. Neurophysiol. *88*, 2530–2546.

Chalk, M., Gutkin, B., and Denève, S. (2016). Neural oscillations as a signature of efficient coding in the presence of synaptic delays. Elife *5*, e13824.

Chalk, M., Herrero, J.L., Gieselmann, M.A., Delicato, L.S., Gotthardt, S., and Thiele, A. (2010). Attention reduces stimulus-driven gamma frequency oscillations and spike field coherence in V1. Neuron *66*, 114–125.

Chao, Z.C., Takaura, K., Wang, L., Fujii, N., and Dehaene, S. (2018). Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. Neuron *100*, 1252–1266.e3.

Chollet, F. (2015). Keras. GitHub. https://github.com/fchollet/keras.

Coen-Cagli, R., Dayan, P., and Schwartz, O. (2009). Statistical models of linear and non-linear contextual interactions in early visual processing. In Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference (Curran Associates, Inc), pp. 369–377, ISBN: 9781615679119. https://proceedings.neurips.cc/paper/2009/file/be3159ad04564bfb90db9e32851ebf9c-Paper.pdf.

Coen-Cagli, R., Dayan, P., and Schwartz, O. (2012). Cortical surround interactions and perceptual salience via natural scene statistics. PLoS Comput. Biol. *8*, e1002405.

Coen-Cagli, R., Dayan, P., and Schwartz, O. (2016). MATLAB tools for building mixture of Gaussian scale mixture (MGSM) models, and perform inference and learning (CRCNS.org). https://doi.org/10.6080/K0JM27JZ.

Coen-Cagli, R., Kohn, A., and Schwartz, O. (2015). Flexible gating of contextual influences in natural vision. Nat. Neurosci. *18*, 1648–1655.

Criminisi, A., Pérez, P., and Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process *13*, 1200–1212.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D.D., and DiCarlo, J.J. (2020). Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. bioRxiv. bioRxiv. https://doi.org/10.1101/2020.06.16.154542.

Das, A., and Ray, S. (2018). Effect of stimulus contrast and visual attention on spike-gamma phase relationship in macaque primary visual cortex. Front. Comp. Neurosci. *12*, 66.

de Lange, F.P., Heilbron, M., and Kok, P. (2018). How do expectations shape perception? Trends Cogn. Sci. *22*, 764–779.

Dosovitskiy, A., and Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. In Advances in Neural Information Processing Systems, pp. 658–666.

Einhäuser, W., and König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? Eur. J. Neurosci. *17*, 1089–1097.

Engel, A.K., Kreiter, A.K., König, P., and Singer, W. (1991). Synchronization of oscillatory neuronal responses between striate and extrastriate visual cortical areas of the cat. Proc. Natl. Acad. Sci. USA *88*, 6048–6052.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., et al. (2019). U-net: deep learning for cell counting, detection, and morphometry. Nat. Methods *16*, 67–70.

Falkner, S., Klein, A., and Hutter, F. (2018). BOHB: robust and efficient hyperparameter optimization at scale. In Proceedings of the 35th International Conference on Machine Learning, J. Dy and A. Krause, eds. (Stockholmsmässan), pp. 1437–1446.

Ferro, D., van Kempen, J., Boyd, M., Panzeri, S., and Thiele, A. (2021). Directed information exchange between cortical layers in macaque V1 and V4 and its modulation by selective attention. Proc. Natl. Acad. Sci. USA *118*, e2022097118.

Fries, P. (2015). Rhythm for cognition: communication through coherence. Neuron *88*, 220–235.

Fries, P., and Maris, E. (2021). What to do if n is two? arXiv, arXiv:2106.14562.

Fries, P., Reynolds, J.H., Rorie, A.E., and Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. Science *291*, 1560–1563.

Friston, K. (Nov. 2008). Hierarchical models in the brain. PLoS Comp. Biol. *4*, e1000211.

Friston, K. (2010). The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. *11*, 127–138.

Galuske, R.A.W., Munk, M.H.J., and Singer, W. (2019). Relation between gamma oscillations and neuronal plasticity in the visual cortex. Proc. Natl. Acad. Sci. USA *116*, 23317–23325.

Gatys, L.A., Ecker, A.S., and Bethge, M. (2015). A neural algorithm of artistic style. arXiv, arXiv:1508.06576.

Ghodrati, M., Morris, A.P., and Price, N.S.C. (2015). The (un)suitability of modern liquid crystal displays (LCDs) for vision research. Front. Psychol. *6*, 303.

Gieselmann, M.A., and Thiele, A. (2008). Comparison of spatial integration and surround suppression characteristics in spiking activity and the local field potential in macaque V1. Eur. J. Neurosci. *28*, 447–459.

Gilbert, C.D., and Wiesel, T.N. (1983). Clustered intrinsic connections in cat visual cortex. J. Neurosci. *3*, 1116–1133.

Gray, C.M., König, P., Engel, A.K., and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature *338*, 334–337.

Gray, C.M., and McCormick, D.A. (1996). Chattering cells: superficial pyramidal neurons contributing to the generation of synchronous oscillations in the visual cortex. Science *274*, 109–113.

Gregoriou, G.G., Gotts, S.J., Zhou, H., and Desimone, R. (2009). High-frequency, long-range coupling between prefrontal and visual cortex during attention. Science *324*, 1207–1210.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. Cogn. Sci. *11*, 23–63.

Grothe, I., Neitzel, S.D., Mandon, S., and Kreiter, A.K. (2012). Switching neuronal inputs by differential modulations of gamma-band phase-coherence. J. Neurosci. *32*, 16172–16180.

Hadjipapas, A., Lowet, E., Roberts, M.J., Peter, A., and De Weerd, P. (2015). Parametric variation of gamma frequency and power with luminance contrast: a comparative study of human MEG and monkey LFP and spike responses. NeuroImage *112*, 327–340.

Haefner, R.M., Berkes, P., and Fiser, J. (2016). Perceptual decision-making as probabilistic inference by neural sampling. Neuron *90*, 649–660.

Hartmann, T.S., Raja, S., Lomber, S.G., and Born, R.T. (2019). Cortico-cortical feedback from V2 exerts a powerful influence over the visually evoked local field potential and associated spike timing in v1. bioRxiv. bioRxiv. https://doi.org/10.1101/792010.

Hasenstaub, A., Shu, Y., Haider, B., Kraushaar, U., Duque, A., and McCormick, D.A. (2005). Inhibitory postsynaptic potentials carry synchronized frequency information in active cortical networks. Neuron *47*, 423–435.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In The IEEE International Conference on Computer Vision (ICCV).

Heeger, D.J. (2017). Theory of cortical function. Proc. Natl. Acad. Sci. USA *114*, 1773–1782.

Henrie, J.A., and Shapley, R. (2005). LFP power spectra in V1 cortex: the graded effect of stimulus contrast. J. Neurophysiol. *94*, 479–490.

Hermes, D., Miller, K.J., Wandell, B.A., and Winawer, J. (2015). Stimulus dependence of gamma oscillations in human visual cortex. Cereb. Cortex *25*, 2951–2959.

Hermes, D., Petridou, N., Kay, K.N., and Winawer, J. (2019). An image-computable model for the stimulus selectivity of gamma oscillations. Elife *8*, e47035.

Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.

Hutter, F., Hoos, H., and Leyton-Brown, K. (2014). An efficient approach for assessing hyperparameter importance. In Proceedings of International Conference on Machine Learning 2014 (ICML 2014)), pp. 754–762.

Iglovikov, V., and Shvets, A. (2018). Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation, arXiv, arXiv:1801.05746.

Iqbal, H. (2018). Plotneuralnet. Zenodo. https://zenodo.org/record/2526396.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision (Springer), pp. 694–711.

Kapadia, M.K., Westheimer, G., and Gilbert, C.D. (1999). Dynamics of spatial summation in primary visual cortex of alert monkeys. Proc. Natl. Acad. Sci. USA *96*, 12073–12078.

Keller, A.J., Roth, M.M., and Scanziani, M. (2020). Feedback generates a second receptive field in neurons of the visual cortex. Nature *582*, 545–549.

Keller, G.B., and Mrsic-Flogel, T.D. (2018). Predictive processing: a canonical cortical computation. Neuron *100*, 424–435.

Kingma, D.P., and Ba, J. (2014). Adam: a method for stochastic optimization, arXiv, arXiv:1412.6980.

Kirchberger, L., Mukherjee, S., Schnabel, U.H., van Beest, E.H., Barsegyan, A., Levelt, C.N., Heimel, J.A., Lorteije, J.A.M., van der Togt, C., Self, M.W., et al. (2021). The essential role of recurrent processing for figure-ground perception in mice. Sci. Adv. *7*, eabe1833.

König, P., Engel, A.K., and Singer, W. (1996). Integrator or coincidence detector? The role of the cortical neuron revisited. Trends Neurosci *19*, 130–137.

Lamme, V.A., and Spekreijse, H. (1998). Neuronal synchrony does not represent texture segregation. Nature *396*, 362–366.

LeCun, Y. (2019). 1.1 Deep learning hardware: past, present, and future. In 2019 IEEE International Solid-State Circuits Conference (ISSCC) (IEEE)), pp. 12–19.

Lee, J., Cho, S., and Beack, S.-K. (2018). Context-adaptive entropy model for end-to-end optimized image compression, arXiv, arXiv:1809.10452.

Legatt, A.D., Arezzo, J., and Vaughan, H.G. (1980). Averaged multiple unit activity as an estimate of phasic changes in local neuronal activity: effects of volume-conducted potentials. J. Neurosci. Methods *2*, 203–217.

Li, X., Chen, Y., Lashgari, R., Bereshpolova, Y., Swadlow, H.A., Lee, B.B., and Alonso, J.M. (2015). Mixing of chromatic and luminance retinal signals in primate area V1. Cereb. Cortex *25*, 1920–1937.

Li, Z. (2002). A saliency map in primary visual cortex. Trends Cogn. Sci. *6*, 9–16.

Lima, B., Singer, W., Chen, N.H., and Neuenschwander, S. (2010). Synchronization dynamics in response to plaid stimuli in monkey V1. Cereb. Cortex *20*, 1556–1573.

Lindauer, M., Eggensperger, K., Feurer, M., Biedenkapp, A., Marben, J., Müller, P., and Hutter, F. (2019). Boah: a tool suite for multi-fidelity bayesian optimization and analysis of hyperparameters. arXiv, arXiv:1908.06756.

Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., and Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV)), pp. 89–105.

Lowet, E., Roberts, M.J., Peter, A., Gips, B., and De Weerd, P. (2017). A quantitative theory of gamma synchronization in macaque V1. Elife *6*, e26642.

Marques, T., Schrimpf, M., and DiCarlo, J.J. (2021). Multi-scale hierarchical neural network models that bridge from single neurons in the primate primary visual cortex to object recognition behavior. bioRxiv. bioRxiv. https://doi.org/10.1101/2021.03.01.433495.

Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. Linux J *239*, 2.

Mordvintsev, A., Pezzotti, N., Schubert, L., and Olah, C. (2018). Differentiable image parameterizations (Distill). https://distill.pub/2018/differentiable-parameterizations.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol. Cybern. *66*, 241–251.

Odena, A., Dumoulin, V., and Olah, C. (2016). Deconvolution and checkerboard artifacts (Distill). http://distill.pub/2016/deconv-checkerboard.

Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization (Distill). https://distill.pub/2017/feature-visualization.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. (2018). The building blocks of interpretability (Distill). https://distill.pub/2018/building-blocks/.

Onorato, I., Neuenschwander, S., Hoy, J., Lima, B., Rocha, K.-S., Broggini, A.C., Uran, C., Spyropoulos, G., Klon-Lipok, J., Womelsdorf, T., et al. (2020). A distinct class of bursting neurons with strong gamma synchronization and stimulus selectivity in monkey V1. Neuron *105*, 180–197, e5.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. *2011*, 156869.

Palanca, B.J., and DeAngelis, G.C. (2005). Does neuronal synchrony underlie visual feature grouping? Neuron *46*, 333–346.

Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., and Giró-i-Nieto, X. (2017). Salgan: visual saliency prediction with generative adversarial networks, arXiv, arXiv:1701.01081.

Pesaran, B., Vinck, M., Einevoll, G.T., Sirota, A., Fries, P., Siegel, M., Truccolo, W., Schroeder, C.E., and Srinivasan, R. (2018). Investigating large-scale brain

dynamics using field potential recordings: analysis and interpretation. Nat. Neurosci. *21*, 903–919.

Peter, A., Uran, C., Klon-Lipok, J., Roese, R., Van Stijn, S., Barnes, W., Dowdall, J.R., Singer, W., Fries, P., and Vinck, M. (2019). Surface color and predictability determine contextual modulation of V1 firing and gamma oscillations. Elife *8*, e42101.

Pike, F.G., Goddard, R.S., Suckling, J.M., Ganter, P., Kasthuri, N., and Paulsen, O. (2000). Distinct frequency preferences of different types of rat hippocampal neurones in response to oscillatory input currents. J. Physiol. *529*, 205–213.

Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.

Ray, S., and Maunsell, J.H. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biol *9*, e1000610.

Ray, S., and Maunsell, J.H. (2015). Do gamma oscillations play a role in cerebral cortex? Trends Cogn. Sci. *19*, 78–85.

Ray, S., and Maunsell, J.H.R. (2010). Differences in gamma frequencies across visual cortex restrict their possible use in computation. Neuron *67*, 885–896.

Ringach, D.L., Shapley, R.M., and Hawken, M.J. (2002). Orientation selectivity in macaque V1: diversity and laminar dependence. J. Neurosci. *22*, 5639–5651.

Roberts, M.J., Lowet, E., Brunet, N.M., Ter Wal, M., Tiesinga, P., Fries, P., and De Weerd, P. (2013). Robust gamma coherence between macaque V1 and V2 by dynamic frequency matching. Neuron *78*, 523–536.

Rockland, K.S., and Lund, J.S. (1983). Intrinsic laminar lattice connections in primate visual cortex. J. Comp. Neurol. *216*, 303–318.

Roelfsema, P.R., Lamme, V.A., and Spekreijse, H. (2004). Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. Nat. Neurosci. *7*, 982–991.

Rohenkohl, G., Bosman, C.A., and Fries, P. (2018). Gamma synchronization between V1 and V4 improves behavioral performance. Neuron *100*, 953–963.e3.

Rols, G., Tallon-Baudry, C., Girard, P., Bertrand, O., and Bullier, J. (2001). Cortical mapping of gamma oscillations in areas V1 and V4 of the macaque monkey. Vis. Neurosci. *18*, 527–540.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (Springer), pp. 234–241.

Saleem, A.B., Lien, A.D., Krumin, M., Haider, B., Rosón, M.R., Ayaz, A., Reinhold, K., Busse, L., Carandini, M., and Harris, K.D. (Jan. 2017). Subcortical source and modulation of the narrowband gamma oscillation in mouse visual cortex. Neuron *93*, 315–322.

Sceniak, M.P., Ringach, D.L., Hawken, M.J., and Shapley, R. (1999). Contrast's effect on spatial summation by macaque V1 neurons. Nat. Neurosci. *2*, 733–739.

Schmid, M.C., Schmiedt, J.T., Peters, A.J., Saunders, R.C., Maier, A., and Leopold, D.A. (2013). Motion-sensitive responses in visual area V4 in the absence of primary visual cortex. J. Neurosci. *33*, 18740–18745.

Schneider, M., Broggini, A.C., Dann, B., Tzanou, A., Uran, C., Sheshadri, S., Scherberger, H., and Vinck, M. (2021). A mechanism for inter-areal coherence through communication based on connectivity and oscillatory power. Neuron *109*, 4050–4067.e12.

Schomburg, E.W., Fernández-Ruiz, A., Mizuseki, K., Berényi, A., Anastassiou, C.A., Koch, C., and Buzsáki, G. (2014). Theta phase segregation of input-specific gamma patterns in entorhinal-hippocampal networks. Neuron *84*, 470–485.

Schroeder, C.E., Mehta, A.D., and Givre, S.J. (1998). A spatiotemporal profile of visual system activation revealed by current source density analysis in the awake macaque. Cereb. Cortex *8*, 575–592.

Sejnowski, T.J., and Paulsen, O. (2006). Network oscillations: emerging computational principles. J. Neurosci. *26*, 1673–1676.

Self, M.W., van Kerkoerle, T., Supèr, H., and Roelfsema, P.R. (2013). Distinct roles of the cortical layers of area V1 in figure-ground segregation. Curr. Biol. *23*, 2121–2129.

Shadlen, M.N., and Movshon, J.A. (1999). Synchrony unbound: a critical evaluation of the temporal binding hypothesis. Neuron *24* (67–77), 111.

Shirhatti, V., and Ray, S. (2018). Long-wavelength (reddish) hues induce unusually large gamma oscillations in the primate primary visual cortex. Proc. Natl. Acad. Sci. USA *115*, 4489–4494.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition, arXiv, arXiv:1409.1556.

Singer, W. (1999). Neuronal synchrony: a versatile code for the definition of relations? Neuron *24* (49–65), 111–125.

Singer, W. (2021). Recurrent dynamics in the cerebral cortex: integration of sensory evidence with stored knowledge. Proc. Natl. Acad. Sci. USA *118*, e2101043118.

Sjöström, P.J., Turrigiano, G.G., and Nelson, S.B. (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. Neuron *32*, 1149–1164.

Spyropoulos, G., Dowdall, J.R., Schölvinck, M.L., Bosman, C.A., Lima, B., Peter, A., Onorato, I., Klon-Lipok, J., Roese, R., Neuenschwander, S., et al. (2020). Spontaneous variability in gamma dynamics described by a linear harmonic oscillator driven by noise. bioRxiv. bioRxiv. https://doi.org/10.1101/793729.

Srinivasan, M.V., Laughlin, S.B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. Proc. R. Soc. Lond. B Biol. Sci. *216*, 427–459.

Stopfer, M., Bhagavan, S., Smith, B.H., and Laurent, G. (1997). Impaired odour discrimination on desynchronization of odour-encoding neural assemblies. Nature *390*, 70–74.

Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2015). Yfcc100m: the new data in multimedia research. arXiv, arXiv:1503.01817.

Torralba, A., and Oliva, A. (2003). Statistics of natural image categories. Network *14*, 391–412.

Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., and Geiger, A. (2017). Sparsity invariant CNNs. In 2017 International Conference on 3D Vision (3DV) (IEEE)), pp. 11–20.

Van Der Walt, S., Colbert, S.C., and Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. Comput. Sci. Eng. *13*, 22–30.

van Kerkoerle, T., Self, M.W., Dagnino, B., Gariel-Mathis, M.-A., Poort, J., van der Togt, C., and Roelfsema, P.R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. Proc. Natl. Acad. Sci. USA *111*, 14332–14341.

Veit, J., Hakim, R., Jadi, M.P., Sejnowski, T.J., and Adesnik, H. (2017). Cortical gamma band synchronization through somatostatin interneurons. Nat. Neurosci. *20*, 951–959.

Vinck, M., and Bosman, C.A. (2016). More gamma more predictions: gamma-synchronization as a key mechanism for efficient integration of classical receptive field inputs with surround predictions. Front. Syst. Neurosci. *10*, 35.

Vinck, M., Lima, B., Womelsdorf, T., Oostenveld, R., Singer, W., Neuenschwander, S., and Fries, P. (2010). Gamma-phase shifting in awake monkey visual cortex. J. Neurosci. *30*, 1250–1257.

Vinck, M., Womelsdorf, T., and Fries, P. (2013). Gamma-band synchronization and information transmission. In Principles of Neural Coding, R. Quiroga-Quian and S. Panzeri, eds. (CRC Press).

Vinje, W.E., and Gallant, J.L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. Science *287*, 1273–1276.

Von Helmholtz, H. (1867). Handbuch der physiologischen Optik (Voss) *9*.

Wachtler, T., Sejnowski, T.J., and Albright, T.D. (2003). Representation of color stimuli in awake macaque primary visual cortex. Neuron *37*, 681–691.

Wagatsuma, N., Hidaka, A., and Tamura, H. (2021). Correspondence between monkey visual cortices and layers of a saliency map model based on a deep convolutional neural network for representations of natural images. eNeuro *8*. https://doi.org/10.1523/ENEURO.0200-20.2020.

Wang, P., and Nikolić, D. (2011). An LCD monitor with sufficiently precise timing for research in vision. Front. Hum. Neurosci. *5*, 85.

Wang, Z., Bovik, A.C., Sheikh, H.R., and Simoncelli, E.P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. *13*, 600–612.

Wespatat, V., Tennigkeit, F., and Singer, W. (2004). Phase sensitivity of synaptic modifications in oscillating cells of rat visual cortex. J. Neurosci. *24*, 9067–9075.

Xing, D., Yeh, C.-I., Burns, S., and Shapley, R.M. (2012). Laminar analysis of visually evoked activity in the primary visual cortex. Proc. Natl. Acad. Sci. USA *109*, 13871–13876.

Zandvakili, A., and Kohn, A. (2015). Coordinated neuronal activity enhances corticocortical communication. Neuron *87*, 827–839.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595.

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Organisms/strains		
Rhesus macaque (Macaca mulatta)	Medical Research Council Centre for Macaques, Porton Down, Salisbury, SP4 0JQ	NA
Software and algorithms		
MATLAB version 2020a	Mathworks	https://www.mathworks.com/
FieldTrip	Oostenveld et al., 2011	https://www.fieldtriptoolbox.org
Python version 3.6	Python Software Foundation	https://www.python.org
Tensorflow version 1.14	tensorflow	https://www.tensorflow.org
Keras version 2.2.4	keras	https://keras.io
Numpy version 1.19.5	numpy	https://numpy.org
h5py version 2.9.0	h5py	https://www.h5py.org
Scikit-image version 0.17.2	Scikit-image	https://scikit-image.org
U-Net-Pred	This manuscript	Zenodo Code: https://doi.org/10.5281/zenodo.5794076
Gamma-net	This manuscript	Zenodo Code: https://doi.org/10.5281/zenodo.5794082

RESOURCE AVAILABILITY

Lead Contact

- The lead contact of this study is Martin Vinck.
- Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Martin Vinck (martin.vinck@esi-frankfurt.de).

Materials Availability

- This study did not generate new unique reagents.

Data and Code Availability

- The data that support the findings of this study are available from the lead contact M.V.
- Neural network visualizations were made using PlotNeuralNet (Iqbal, 2018).
- Software and packages used for the analysis include FieldTrip (Oostenveld et al., 2011), matplotlib (Hunter, 2007), numpy (Van Der Walt et al., 2011), docker (Merkel, 2014), tensorflow v1.14 (Abadi et al., 2015), and KERAS (Chollet, 2015).
- Custom software developed are made available on the github repositories including small demos; for Figures 1, 2, 3, and 4 at (https://uranc.github.io/U-Net-Pred/) and Figure S5 at (https://uranc.github.io/gamma-net/).
- DOIs and Zenodo links are listed in the key resources table. All other used software is referenced in the text when used and available online.
- 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

All procedures complied with the German and European regulations for the protection of animals and were approved by the regional authority (Regierungspräsidium Darmstadt). Three adult macaque monkeys (Macaca mulatta) were used in this study. Monkey I was female, 14 years old, and 8 kilograms. Monkey A was male, 10 years old, and 14 kilograms. Monkey H was male, 11 years old, and 17 kilograms.

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

METHOD DETAILS

Surgical procedures

Implantations were made in the left hemisphere of V1. In Monkey H, we implanted a Utah array with 64 microelectrodes (inter-electrode distance 400 μm, half of them with a length of 1 mm and half with a length of 0.6 mm, Blackrock Microsystems), and inserted a reference wire under the dura towards parietal cortex. In monkeys A and I, we implanted a semi-chronic microelectrode array Microdrive (SC32-1, Gray Matter Research), containing 32 independently movable tungsten electrodes (inter-electrode distance of 1.5 mm); here, the microdrive chamber was used as the reference. Note that no histological verification of layers/depths was performed, because the animals are still alive. We estimate that our recordings mainly sample activity from layers 2-4, because the vast majority of recording sites do not show the typical inversion of the first deflection of the event-related potential as is found in the deep layers (Li et al., 2015; Schroeder et al., 1998), this is further supported by the electrode lengths in monkey H. Further details on surgical procedures can be found in Peter et al. (2019).

Behavioral task

All monkeys were trained on a fixation task. The animals were positioned 83 (monkey H) or 64 cm (monkeys A, I) in front of a 22 inch 120 Hz LCD monitor (Samsung 2233RZ, Ghodrati et al. [2015]; Wang and Nikolić, 2011). Monkeys self-initiated trials by fixating on a small fixation spot, which was presented at the center of the screen, and had to maintain fixation during the entire trial. Trials during which the eye position deviated from the fixation spot by more than 0.8 (monkey H) or 1.44 visual deg (monkey I and A) radius were aborted. Note however that eye movements were generally constrained within 1 visual deg, with the majority of eye movements within 0.5 visual deg (Figure S2Ei). Correct trials were rewarded with diluted fruit juice. For further details on the task see Peter et al. (2019).

Recordings

Recordings were made with a Tucker Davis Technologies (TDT) system. Data were filtered between 0.35 and 7500 Hz (3 dB filter cutoffs) and digitized at 24.4 kHz (TDT PZ2 preamplifier). Stimulus onsets were recorded with a custom-made photodiode. Eye movements and pupil size were recorded at 1000 Hz using an Eyelink 1000 system with infrared illumination. For further details see Peter et al. (2019).

Visual stimulation

Image selection

Natural images were acquired from the Yahoo Flickr Creative Commons 100 Million (YFCC100M) Dataset (Thomee et al., 2015). The resolution of these images was high enough to match the resolution of the LCD monitor given the stimulus size (see below). Images were included if user tags included any of the following: Animal, building, closeup, flower, house, indoor, landscape, natural, object, texture, tool, toy, tree. Any of the following tags led to exclusion: Blur, blurry, bokeh, digital, art, artwork, artist, text, writing, drawing, painting, cartoon, graphic, graphic+design, illustration, logo, desktop. A set of 340 images was selected as stimulus set for the monkeys. Note that from the same Yahoo Flickr database, a training and validation set was selected for the deep neural network training (see further below); this training set was different from the images presented to the monkey. All images were converted to grayscale, except for a subset of sessions in which we presented the images also in color. The 340 selected images had to fulfill the following criteria, which are comparable to the ones used in Coen-Cagli et al. (2015): (1) A mean RGB value between 40 and 200. (2) An average luminance-contrast, measured as root-mean-squared (RMS) contrast, above 0.2. (3) A centroid of spatial frequency (defined further below in the Image Statistics section) greater than 0.5 dva (degrees of visual angle). These criteria excluded images that were excessively bright, dark, or spatially uniform. However, note that we studied V1 responses to uniform grayscale surfaces in Peter et al. (2019).

Image standardization

We cropped each image to 600x600 pixels, and applied two further transformations to the images, similar to Coen-Cagli et al. (2015): (1) We set the global luminance-contrast for each image to 0.6, by using a sigmoid projection of pixel values. (2) We rescaled and centered the images to have a mean RGB value of 128, equal to the background. For stimulus presentation, we approximately centered the images on the cluster of recorded V1 RFs. Stimuli had a width of 11.5 dva (monkey H, stimulus centered horizontally at x=+2 and vertically at y=-3 dva from fixation) or 15 dva (monkey A x=+3.55 and y=-0.12, monkey I x=+3.45 and y=-0.02).

Further image selection criteria

Stimulus sets for a given recording session consisted of 20 images, each presented for a total of 10-20 trials. Each session contained a subset of ten images where at least part of the image had a high spatial predictability (>0.85, see below). This was done to ensure that a sufficient number of image patches with high predictability were sampled, given the low probability of finding image patches with very high spatial predictability (Figure S3F). These images were found as follows: (1) Around the clusters of RFs, a region of 3 x 3 degrees visual angle was selected (ROI). (2) The ROI was divided into 1 degree patches and the spatial predictability of 9 non-overlapping patch locations were quantified (see section Image Statistics). (3) At least 2 out of 9 patch locations were required to have a predictability value (defined with SSIM) above 0.85. The other 10 images were, in terms of predictability, selected randomly. Note that correlations between spatial predictability and synchronization were also found below 0.85 (Figure 2C), and that for all 20 images,

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

there were typically some recordings sites with RFs corresponding to low predictability. Additionally, all 20 images were required to have, inside the ROI, an average luminance-contrast above 0.15 and a centroid of the spatial frequency below 8 cycles per degree; this prevents aliasing, given the visual acuity of macaques at the recorded eccentricities.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data analysis

Preprocessing

We analyzed only correctly performed trials. We downsampled the LFPs to $\approx$1.02kHz using Matlab's decimate.m function. Line noise was removed using a two-pass 4th order Butterworth bandstop filters between 49.9-50.1, 99.7-100.3 and 149.5-150.5 Hz. Similar to previous studies (Schmid et al., 2013; Self et al., 2013; Xing et al., 2012; Legatt et al., 1980; Peter et al., 2019), we computed MU (Multi Unit) signals from the broadband signal, by bandpass filtering between 300 and 6000 Hz (4th order butterworth), rectification, and applying low-pass filtering and downsampling in the same way as for the LFPs. For the calculation of rate modulations, this MU signal was smoothed with a Gaussian kernel with an SD of 20 ms. For further details see Peter et al. (2019).

RF estimation

RF were mapped with moving bar stimuli (spanning the entire monitor). Moving bars (width 0.1 deg, speed 10-17 deg/s) were presented with 8-16 different orientations. MU responses were computed as a function of stimulus position, after correcting for response latency. MU responses for each movement direction were then fitted by a Gaussian function. We used this fit to extract the 10th percentile and the 90th percentile. Across the 16 directions, this yielded 32 data points, which were fit with an ellipse. This ellipse was defined as that MU's RF. The RF size was defined as the diameter based on $2 \times \sqrt{a \times b}$, where a and b are the major and minor radius. The preferred stimulus orientation was also computed using these bar stimuli (and was highly consistent with orientation tuning based on static gratings).

Electrode selection

We included all electrodes for analysis that met the following criteria: (1) The MU showed a response to RF stimulation that was at least two SD above stimulation outside the RF. (2) The MU response during the response period (0.05-0.15 s post stimulus onset) of at least one stimulus condition in the session was at least 2 SD above the corresponding baseline (-0.4-0 s before stimulus onset).

Estimation of LFP power spectra

For the experiment in Figures 7 and S7 (masked paradigm), the baseline period was the last 400 ms before stimulus onset, and each stimulation period yielded an epoch of 400 ms (0.2-0.6 s period). For the rest of the figures, the baseline period was the last 500 ms before stimulus onset, and each stimulation period yielded two non-overlapping epochs of 500 ms (0.2-1.2 s period). For monkey H, we made use of the constant spacing of neighboring electrodes in the array to improve power spectral estimation: Power spectra were approximated by the complex mean of the cross-spectra of a channel with its two same-depth nearest neighbors. This reduced uncorrelated high-frequency fluctuations due to spiking, which can affect the $1/f$ slope of LFP spectra; qualitatively similar results were obtained using the unipolar power spectra. LFP epochs were multi-tapered ($\pm$5 Hz smoothing) (Pesaran et al., 2018), Fourier transformed and squared to estimate LFP power spectral densities.

Quantification of LFP gamma-band and beta-band amplitude

In Peter et al. (2019), we developed an algorithm to extract the amplitude of narrow-band gamma-band (or beta-band) oscillations, with several advantages compared to previous methods. Note that qualitatively similar correlations with spatial predictability were obtained by simply using baseline-corrected power in β and γ bands (see Figure S2A), however baseline-corrected power can be skewed by firing rates which results in a stronger correlation with luminance-contrast (Figures S4D and S4E). The algorithm had the following structure:

1. Power spectra were log-transformed and the frequency axis was also sampled in log-spaced units to avoid overweighing the contribution of high-frequency datapoints. All subsequent polynomial fits were performed on the 5-200 Hz range.
2. 1/F^n correction was performed by fitting an exponential to the LFP power spectrum, excluding data points in the range of 10-85 Hz in order to avoid any influence on peak detection. (For a subset of figures, namely Figures S2B, S4D, and S4E, we corrected the power by dividing by the pre-stimulus baseline).
3. To determine the polynomial order, we used a cross validation procedure to prevent overfitting. Polynomials of order 1-20 were fit to ΔP as a function of frequency for a ''training set''. We then evaluated the mean squared error using the same polynomial fit on a ''test set'' for each of the 20 orders. This procedure was then repeated for multiple iterations and we chose the order with the best median performance.
4. On the polynomial fit, local maxima and minima in the β (18-30 Hz) or γ (30-80 Hz) range were identified. The peak frequency was the location of the maximum. The band-width was estimated as twice the distance between the frequency of the maximum (F_{max}) and the frequency of the first local minimum to the left of the maximum (F_{min}), i.e. $b = F_{min} + 2 \cdot (F_{max} - F_{min}) = 2F_{max} - F_{min}$. The amplitude was then assessed from the difference between the value of the polynomial fit at the maximum and the average of the polynomial fit at F_{min} and $2F_{max} - F_{min}$, where F_{min} is the frequency of the nearest local minimum to the left of the maximum (we used the left one, to avoid any influence of spike-bleed-in at higher frequencies).
5. The amplitude was quantified as a fold-change.

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

Rate modulation

Rate modulation was computed as

$$R = (M_{stim} / M_{base}) , \quad \text{(Equation 1)}$$

where M_{stim} and M_{base} represent the MU firing activity in the stimulus and baseline period, respectively. Spike-density functions were normalized in the same way.

Statistics

Error bars or shaded error regions correspond to ± one standard errors of the mean (SEM) across recording sites. Violin plots show the median together with the 25-75 percentiles and the data distribution estimated using Matlab's ksdensity.m function. The statistics used in the rest of the manuscript were as follows:

- In Figures 2C, 4B, and 6C, we show average neural activity (72 recording sites) as a function of spatial predictability or luminance-contrast. We pooled all 72 channels and RF image patches together and then formed non-overlapping bins of 250 RF image patches. We then performed a Spearman-rank correlation across these 250 images.
- In Figures 2D and 4C, we correlated γ-peak amplitude with predictability or luminance-contrast across sessions for each channel separately. We tested whether the average correlation was significantly different from zero across 72 channels by using a two-sided T-test.
- In Figures 4C and 6D, we tested whether different models were significantly different from each other by using 10-fold cross-validation, and a paired T-test across 72 recording sites. The additional regression variables used in Figures 4C were: Center luminance, center spatial frequency and absolute deviation from preferred MU orientation, as well as the luminance and spatial frequency of the 224 × 224 pixel image patch surrounding the RF image patch.
- In Figures 3E and S3E, we performed a two-sided T-test between nature and man-made or object-boundary vs. no boundary categories (N=72 recording sites).
- In Figure 6D, we used a Pearson correlation to quantify the trend across layers.
- In Figure 5B, we compared the strength in correlations between blocks of VGG-16 layers using a pairwise two-sided T-test across N=72 recording sites (early: layer 1-4; middle: layer 5-9; deep: 10-13).
- We repeated 5 experimental sessions twice with 5 different stimulus sets with 100 stimuli. Using this we estimated the variance of the estimated means (which mainly results due to limited number of trials) and corrected for this variance.

Note that statistical inference, as in almost all electrophysiological studies, can be argued to be limited to the analyzed sample (Fries and Maris, 2021).

Identification of object boundaries and man-made vs. nature pictures

To automatically identify object boundaries, we used the Mask-R convolutional neural network. Using this network, we determined if there is a detected object contour that intersects with the MU's RFs. We distinguished man-made vs. nature pictures as follows: Pictures were categorized as man-made if man-made structure (e.g. buildings) was within the 4 × 4 degree image region centered on the RFs of the recording array (for examples see Figure S3C). Images with both nature and man-made content were not considered for this comparison.

Deep Neural Networks methods

We used DNNs for several purposes. In Figures 1, 2, 3, 4, 6, and 7 we trained a network to predict missing parts of an image to obtain several measures of spatial predictability. We will first describe the stimuli used to train this network, its architecture and training procedures. Following this, we describe other networks in which we directly predicted neural activity from the image.

Preprocessing of training stimuli

We resized the images from the Yahoo Flickr dataset (see Section "Visual Stimulation") to 300 × 300 pixels. To improve robustness and generalization, we applied standard data augmentation of the images for the training, consisting of several operations from the Tensorflow image module (Abadi et al., 2015). Each operation had a 50% chance of being applied for a given image: brightness (max_delta=0.1), contrast (= [0, 1]), hue (max_delta = 0.1), saturation (= [0, 1]), convert to black & white, horizontal flip. Resulting images were then randomly cropped to 224 × 224 pixels. The mean RGB value [123.68, 116.779, 103.939] of ImageNet was subtracted from each image, in order to use the VGG-16 network for the initial layers of the network (Simonyan and Zisserman, 2014).

Mask generation

In order to make a network that can robustly predict missing inputs, we trained the network with binary masks that either had occluders or missing pixels (low signal-to-noise ratio). Binary masks were randomly selected from 3 types: elliptic, rectangular, or salt and pepper noise. Rectangular and elliptic mask types consisted of 2-3 missing regions that were randomly selected to be 20 × 20 pixels to 80 × 80 pixels (0.5-2 degrees) in area. One side of the rectangle (or axis for the ellipse) was randomly picked between 20 to 80 pixels. Finally, random rotations were applied to each missing region. Salt & Pepper noise had a sparsity ratio of at least 20%. Images including the mask were randomly cropped to size 224 × 224, and horizontally flipped with a 50% chance.

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

Architecture of deep neural network for inpainting

For the inpainting we relied on Deep Neural Networks. Note that qualitatively similar correlations (however of smaller magnitude) were obtained when generating image predictions based on an algorithm that does not use deep neural networks (Criminisi et al., 2004), demonstrating the generality of our approach (Figure S2D). The neural network architecture was based on the U-Net architecture (Falk et al., 2019; Ronneberger et al., 2015), with the following modifications: For initialization, the encoder part of U-Net was replaced by all the convolutional and pooling layers of the VGG-16 network, using the Keras implementation (Simonyan and Zisserman, 2014; Chollet, 2015). Transfer learning using VGG-16 has been previously used in image segmentation (Iglovikov and Shvets, 2018), image reconstruction (Uhrig et al., 2017), style transfer (Gatys et al., 2015), and image inpainting (Liu et al., 2018). The resulting network architecture consisted of five blocks, each of which had two or three convolutional layers (3 × 3) with ReLU (rectified linear) activation functions, followed by a max-pooling (2 × 2) operation. The decoder consisted of five blocks, each with a nearest-neighbor upsampling layer (2 × 2), followed by two convolutional layers. The output layer was a convolutional layer with, as is conventional, a linear activation function.

Partial Convolution

All convolution operations in the network, including the VGG-16 network, were implemented as partial convolutions. Partial convolution has been introduced with the sparsity-invariant convolutional network where the input to each convolution is paired with a binary mask indicating which pixels are observable or missing, respectively (Uhrig et al., 2017). Partial observability of the inputs during the training makes the network robust to input sparsity, regardless of the task of the network. We implemented a modified version with mask updates per network operation, as described in Liu et al. (2018). The idea of partial convolution is that the missing region is gradually filled, and that the filled-in information is used for filling in the rest of the missing pixels in an iterative way.

Loss function for training

Convolutional neural network (CNN) activations have been previously used as a basis for perceptual similarity metrics instead of traditional measures such as SSIM (Wang et al., 2004) or L2-norm (Dosovitskiy and Brox, 2016; Gatys et al., 2015; Zhang et al., 2018). Even though these perceptual loss functions based on CNNs match human perception better, image generation based on those loss functions can suffer from high frequency artifacts (Olah et al., 2017, 2018; Johnson et al., 2016). To minimize high-frequency artifacts, we therefore implemented the reconstruction and content loss functions in the Fourier domain, similar to decorrelated image parameterization described in Mordvintsev et al. (2018) and Odena et al. (2016).

Total loss consisted of reconstruction, content, and style losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reconstruction}} + p_{\text{content}} \cdot \mathcal{L}_{\text{content}} + p_{\text{style}} \cdot \mathcal{L}_{\text{style}}$$

Here, p_{style} and p_{content} are hyperparameters (see further below). Reconstruction loss consisted of three terms: (i) The difference between the amplitudes of Fourier transforms; (ii) the log difference between the Fourier amplitudes; (iii) the phase similarity between the predicted and the original image, which has been previously applied to auditory signal synthesis in the Fourier domain (Arık et al., 2018):

$$\mathcal{L}_{\text{reconstruction}} = \mathrm{F}(y, \hat{y}) \; ;$$

$$\mathrm{F}(y, \hat{y}) = \ell_1(||\mathfrak{F}(y)|| - ||\mathfrak{F}(\hat{y})||)$$

$$+ p_{\log} \cdot \ell_1(\ln||\mathfrak{F}(y)|| - \ln||\mathfrak{F}(\hat{y})||)$$

$$+ p_{\text{phase}} \cdot \ell_1\big(\, ||\mathfrak{F}(y)|| \cdot ||\mathfrak{F}(\hat{y})|| \,$$

$$- \, ||\mathrm{Re}(\mathfrak{F}(y))|| \cdot ||\mathrm{Re}(\mathfrak{F}(\hat{y}))|| \,$$

$$- \, ||\mathrm{Im}(\mathfrak{F}(y))|| \cdot ||\mathrm{Im}(\mathfrak{F}(\hat{y}))|| \, \big)$$

Here, y and $\hat{y}$ are the original and predicted image. The operator $\mathfrak{F}$ denotes the Discrete Fourier Transform, $||\mathfrak{F}(y)||$ denotes the Fourier magnitude, and ℓ_1 denotes the L1-norm across all frequencies.

Content loss was also defined in the Fourier domain, taking as input the cth AN activation in the λth layer activations of the VGG-16 network:

$$\mathcal{L}_{\text{content}} = \sum_{c} \sum_{\lambda} \mathrm{F}(\varphi_{\lambda,c}(y), \varphi_{\lambda,c}(\hat{y}))$$

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

Finally, style loss is the difference between the Gramian matrices of the Fourier amplitude of the predicted and the original images, where the Gramian matrices contain the covariance matrix across AN activations:

$$\mathcal{L}_{\text{style}} = \sum_{\lambda} \ell_1 \left(\mathrm{G}_{cc'} \left(||\widetilde{\mathcal{F}} \left(\varphi_{\lambda,c} \left(\mathrm{y} \right) \right)|| \right) - \mathrm{G}_{cc'} \left(||\widetilde{\mathcal{F}} \left(\varphi_{\lambda,c} \left(\widehat{\mathrm{y}} \right) \right)|| \right) \right)$$

Averages of layers conv1_2, conv2_2, conv3_3, conv4_3 of VGG-16 network were used for content and style loss.

Training and Hyperparameter Optimization

We initialized the weights of the encoder as VGG-16 model weights that were pre-trained on ImageNet. The remaining weights were initialized using He-initialization (He et al., 2015) and bias terms as 0. The network weights were optimized using the Adam optimizer with a learning rate of $5e^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e^{-7}$, (Kingma and Ba, 2014). All hyperparameters were defined as continuous variables where the search space was [1e-5, 1e3] sampled in logarithmic scale (Lindauer et al., 2019). We optimized the hyperparameters using a method that combines Bayesian optimization with Hyperband as described in Falkner et al. (2018). We ran the optimization for 50 iterations with a minimum budget of 8 and a max budget of 16. For hyperparameter optimization, we used a subset of the image dataset for training (16384 images) and validation (1024 images) of the network. We used SSIM based reconstruction loss as the evaluation metric for the hyperparameter optimization. To analyze the hyperparameter space and the importance of individual hyperparameters, we used fANOVA (Hutter et al., 2014). The importance of a hyperparameter is the fraction of explained variance (mean across 100 repetitions ± SEM) of the validation SSIM-loss across the entire hyperparameter space. The resulting hyperparameters and their importances were:

Name	Value	Importance
p_{style}	0.18	0.13 ± 0.002
$p_{content}$	36.90	0.22 ± 0.002
p_{log}	5.54	0.14 ± 0.002
p_{phase}	3.02e-05	0.29 ± 0.003
p_{lr}	4.48e-05	0.14 ± 0.002

Image Statistics

Spatial Predictability

For a given image, predictability was computed by masking out the central 1 degree patch (similar in size to previous studies studying contextual modulation, Coen-Cagli et al. (2015)). The RF image-patch was then predicted by the inpainting algorithm described above. The to-be-predicted image patch is denoted **I**, a matrix of $N \times N$ pixels.

Structural predictability

Structural predictability was defined as the squared Pearson correlation of two images or average correlation across VGG-16 AN activations. This was defined for each layer λ of the VGG-16 separately:

$$\rho_{\lambda} = \frac{1}{C_{\lambda}} \sum_{c} \frac{\mathrm{cov}(\varphi_{\lambda,c}(y), \varphi_{\lambda,c}(\widehat{y}))}{\sigma_{\varphi_{\lambda,c}(y)} \sigma_{\varphi_{\lambda,c}(\widehat{y})}}$$

Content predictability

Content loss was defined as the L2-norm of the difference between the VGG-16 λ layer AN activations of two images. This was defined for each layer of the VGG-16 separately:

$$L_{\lambda} = \frac{1}{I_{\lambda} J_{\lambda} C_{\lambda}} \sum_{ij,c} \ell_2 \left(\varphi_{\lambda,c}(y_{ij}) - \varphi_{\lambda,c}(\widehat{y}_{ij}) \right)$$

Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS is computed as the average content loss of layers conv1_2, conv2_2, conv3_3, conv4_3 conv5_3. Please note that we used the VGG-16 net-lin model without the learned weights as described in Zhang et al. (2018).

Luminance-contrast

Luminance-contrast was measured as the Root-mean-square (RMS) contrast:

$$\mathrm{RMS} = \sqrt{\frac{1}{N^2} \sum_{ij} (I_{ij} - \bar{I})^2},$$

where $\bar{I}$ is the mean value of I. RMS was defined to range between 0 (minimum) and 1 (maximum) luminance-contrast.

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

Spectral Centroid

The center of mass of the power spectrum is the mean spatial frequency in the image patch, f_k, weighted by the total power, r_k in the frequency bands.

$$R = \frac{\sum_k r_k f_k}{\sum_k r_k}.$$

Orientation

Mean orientation in the RF was computed weighted by the total power in the orientation band [0:5:180] degrees, across spatial frequencies 1.0, 2.8284, 8.0 cycles per degree, similar to the orientation selectivity of the receptive field. We estimated the orientation selectivity of the RF as the weighted circular mean across different orientations as described in Ringach et al. (2002).

$$R = \frac{\sum_k r_k e^{i2\theta_k}}{\sum_k r_k}.$$

R is the population resultant vector, r_k is the peak MU response for the orientation θ_k. Orientation selectivity magnitude is given by $|R|$ and the mean angle is given by $\arg(R)$.

Compressibility

Compressibility was defined as the negative bits per pixel (bpp) which is commonly used to benchmark image compression methods. For image compression, we used a context-adaptive, entropy-based deep neural network model that outperforms the traditional image codecs such as BPG or JPEG, as well as other previous DNN based approaches (Lee et al., 2018). For each RF image patch and its surround (224x224 pixels), we compressed the image using the MS-SSIM optimized model with the quality level set to 5 to get a bits-per-pixel (bpp) measure of the compression. For the same image patches we computed the average predictability across 52x52 pixel non-overlapping sub-regions of the whole patch.

Dimensionality

Dimensionality was determined by first taking the two-dimensional fast Fourier transform of the RF image-patch, taking the rotational average, and ranking the spectral components by magnitude. Dimensionality was then defined as the slope of the resulting spectrum.

Homogeneity

We computed P(homogeneity) using the inference model described in Coen-Cagli et al. (2012) similar to Coen-Cagli et al. (2009, 2015). This model consisted of 72 filters with 4 orientations (0, 45, 90, 135 degrees) and 2 phases (even, odd-symmetric quadrature pairs) each, at 9 locations (center and 8 surround locations circular around the center RF with a radius of 6 pixels). Center and surround RFs had a diameter of 9 pixels and peak spatial frequency of 1/6 cycles/pixel. We trained the model with our natural image dataset, downsampled by a factor of 0.22 to match the model RF size (9 pixels) using the code provided by the authors (Coen-Cagli et al., 2016). P(homogeneity) was computed as 1-P(heterogeneity), where P(heterogeneity) is the average inferred probability of heterogeneity of 4 center units which was the output of the inference model.

Orientation Variance

We filtered RF image patches of 224 x 224 pixels by gabor filters as described in Hermes et al. (2019) using the code provided. Model parameters x,y were set to the RF center of the respective recording site, sigma was set to 1 dva, gain(g) was set to 1 and exponent (n) was set to 2.

Visual Saliency

We used SalGAN, a state-of-the art generative adversarial network trained for visual saliency prediction of the 224 × 224 RF image patch (Pan et al., 2017; Wagatsuma et al., 2021) (see Figure S4I). We quantified salience as the average visual saliency prediction value within the RF center.

Human perceptual similarity

We used the Berkeley Adobe Perceptual Patch Similarity Just Noticeable Distance (JND) dataset to evaluate content and structure losses as human perceptual similarity metrics. In the JND experiments, participants were presented two image patches for 1 second each, with a 250 millisecond gap in between. They were asked if the patches were the same or not. In order to evaluate content and structure perceptual similarity metrics, we used the layer specific VGG-16 AN activations of the original and distorted image as predictors for a logistic regression classifier to predict human perceptual similarity judgments. We calculated the Area Under Curve (AUC) to quantify how well the different VGG-16 layers and similarity metrics (LPIPS, LPIPS structure, SSIM, structure) explain human perceptual similarity judgments.

Predicting neural activity from VGG-16 activations

In Figure 5, we predicted neural activity from the VGG-16 AN activations. If VGG-16 ANs had a smaller RF than 1 dva, then only VGG-16 ANs were used with a RF center within the 1 degree region around the multi-unit's RF center. If VGG-16 ANs had a larger RF than 1 dva, then only VGG-16 ANs were used that fully covered the central 1 degree.

Please cite this article in press as: Uran et al., Predictive coding of natural images by V1 firing rates and rhythmic synchronization, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.01.002

For each location in the VGG-16, we predicted the neural responses from a vector of VGG-16 ANs with different feature selectivities. For this, we used 10% of the stimuli as a test set and used the rest as the training set. We used L1-constrained linear regression and 10-fold cross validation to select the L1-constraint parameter λ. Regression coefficients that best explain the training set were then used to predict the neural signals for the test set. The correlation values (r) were averaged across VGG-16 locations. Receptive field sizes in the VGG-16 are shown in Table S1. We calculated the receptive fields for VGG-16 ANs as described in Araujo et al. (2019). The receptive field sizes of the neural network in the middle layers were comparable to the receptive field sizes in V1, which has been previously described in Cadena et al. (2019).

Prediction from VGG-16 activations: gamma-net

In Figures S5C–S5E, we used a convolutional neural network to predict γ-peaks based on the input image. In total, training data consisted of 31988 image patches, by pooling across monkeys, channels and sessions. We used all image patches from a unique %10 of the stimuli as a test set and used the rest for training. We resized the 224x224 image patches to 84x84 in order to reduce the number of parameters in the network. We used the VGG-16 model from keras applications, with frozen weights pre-trained on ImageNet for transfer learning. The VGG-16 activations were the input to a network that consisted of 2 convolutional layers and a readout layer. We compared predictions from different VGG-16 input layers as input (Figure S5D). Convolutional layers consisted of (3x3) filters with bias, stride (2), valid padding, L1-norm kernel regularization (0.001), leaky ReLU activation (0.1), and dropout (0.5). The final two convolutional layers had 32 and 16 ANs. The readout layer consisted of (4x4) filters with bias and leaky ReLU activation (0.1).