BACKPROPAGATION MISALIGNS WITH THE HIERARCHY OF BRAIN RESPONSES TO IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Backpropagation is the core learning mechanism underlying deep learning. However, whether and how this algorithm is implemented in the brain remains highly debated. In particular, while forward activations of pretrained models reliably map onto the cortical hierarchy of visual processing, it is unknown whether backpropagated gradients exhibit a similar correspondence. Here, we address this question using functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) recordings of human brain responses to natural images. For this, we extend standard encoding analyses of forward activations to map backpropagated gradients onto neural data. Focusing on a recent self-supervised vision model (DINOv3) and reproducing results on eight vision models, we find that backpropagated gradients can reliably predict both fMRI and MEG signals, specifically in higher-level visual cortex and for later latencies. However, the spatial and temporal organization of these backpropagated gradients in the brain diverges from the patterns expected under a biologically plausible backpropagation mechanism: specifically, both the order in which gradients are computed and their spatial organization diverge from the temporal and spatial hierarchies of the human brain. Together, these results suggest that, although deep networks and the brain may share similar representational content, they likely rely on fundamentally different mechanisms to learn those representations.

1 INTRODUCTION

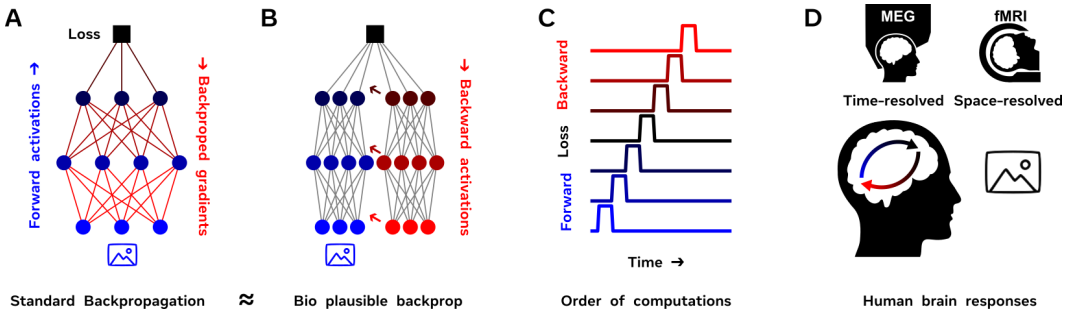


Figure 1: **Tracking Backpropagation in the Brain. A.** The standard backpropagation algorithm first computes a bottom-up hierarchy of activations, then evaluates the loss, and finally computes a top-down hierarchy of gradients to update network weights. **B.** A biologically plausible implementation of backpropagation relies on a top-down neural network with a backward stream of activations (Lillicrap et al., 2020). This framework predicts that gradient-related information should also be reflected in neural activations. **C.** Standard and biologically plausible implementations of backpropagation therefore share a similar computational sequence, yielding testable predictions for neuroscience. **D.** Here, we analyze human brain responses to natural images recorded with MEG and ultra-high-field fMRI (7T), providing high temporal and spatial resolutions. We then evaluate the encoding of forward and backward activations from DINOv3 (Siméoni et al., 2025) and other models to characterize the timing and anatomical localization of these representations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Backpropagation as a learning principle. Backpropagation is a core element of deep learning: it computes, in a top-down manner, the gradient of the loss with respect to each synaptic weight in a potentially very deep hierarchical network, enabling loss minimization during training Rumelhart et al. (1986). Whether the brain implements backpropagation or a functionally equivalent mechanism remains a central and yet unresolved question in computational neuroscience and artificial intelligence Lillicrap et al. (2020); Whittington and Bogacz (2017).

Similar visual representations between AIs and the Brain. The biological plausibility of this algorithm stands in contrast to that of its "forward" counterpart. Indeed, although deep neural networks are not designed to mimic the brain, their forward activations partially converge toward representations similar to those observed in the mammalian visual system Yamins et al. (2014); Khaligh-Razavi and Kriegeskorte (2014); Guclu and van Gerven (2015); Cichy et al. (2014). In particular, bottom-up hierarchy of activations in deep networks have repeatedly been found to align with activations of the cortical hierarchy (Millet et al., 2023; Raugel et al., 2025; Wang et al., 2023)

Remaining challenge. Empirical comparisons between artificial and biological neural networks have almost exclusively focused on forward activations, leaving open the question of how learning – and backpropagation in particular – relates to brain activity. Previous studies have examined neural correlates of prediction errors ((Heilbron et al., 2022) in LLMs). However, to our knowledge, no work has directly tested whether the top-down hierarchy of backpropagated gradients aligns with neural responses.

Biological backpropagation. While the standard implementation of the backpropagation algorithm requires synaptic symmetry, and is thus biologically implausible, Lillicrap et al. (2020) showed that a functionally equivalent but biologically plausible implementation can be computed by approximating backpropagated gradients through a top-down hierarchical neural network. From this perspective, backpropagated gradients can be interpreted as backward activations that convey learning signals across cortical hierarchies.

Predictions. Under this framework, a biological implementation of backpropagation would give rise to three testable signatures. First, backpropagated gradients from the model should be linearly mappable onto the brain responses elicited by the corresponding stimuli. Second, mappings based on forward activations should peak earlier than those based on backpropagated gradients. Third, the spatial and temporal organization of these mappings should reflect a bottom-up progression for forward activations followed by a top-down progression for backpropagated gradients.

Approach To test these predictions, we adapt the standard encoding analyses of forward activations to map backpropagated gradients of a SOTA self-supervised vision model (DINOv3) onto the brain response to a large number of natural images. We combine fMRI (Allen et al., 2022) for spatial resolution and MEG (Hebart et al., 2023) for temporal resolution to evaluate whether these signals align with the computational ordering prescribed by backpropagation. Our work builds on an extensive literature on predictive coding, perceptual learning, and modern theories of cortical plasticity (Rao and Ballard, 1999; Millidge et al., 2022; Sasaki et al., 2009; Furutachi and Hofer, 2026), which demonstrates that, beyond reinforcement learning, the cortex continuously adapts its internal representations from ongoing sensory experience, including during passive viewing, through prediction-error signals – in this framework, perception and learning are inseparable. In both of our datasets, participants were required to recognize and actively process the presented images in recognition or oddball tasks, providing an alert, task-engaged setting in which these learning mechanisms can operate.

Our results confirm the predictions of a similar forward pass as well as a robust mapping between backpropagated gradients and brain activity, explaining neural variance that could not be explained by forward activations alone, specifically in higher-level visual cortex. However, the temporal and spatial organization of this mapping diverges from the top-down organization of the backpropagation algorithm.

2 METHODS

2.1 ENCODING

To evaluate whether, when and where the backpropagated gradient of the loss map onto the brain responses to natural images, we adapt a standard encoding analysis. Following the rationale of (Naselaris et al., 2011), encoding provides an interpretable mapping from model features to neural responses.

Mapping. This linear analysis consists of evaluating whether there exists a linear mapping $W \in \mathbb{R}^{d \times m}$ that reliably predicts m-dimensional brain activity ($Y \in \mathbb{R}^{n \times m}$) given the d-dimensional model activation ($X \in \mathbb{R}^{n \times d}$) in response to n images:

$$\arg\min_{W} \left\{ \|Y - XW\|_2^2 + \lambda \|W\|_2^2 \right\}$$

with λ the ridge regularization parameter. We use linear probes to maintain geometries of compared representations. We use scikit-learn's `RidgeCV` (Pedregosa et al., 2011), 10 logarithmically-spaced regularization λ in between 10^0 and 10^8, and a 5-split cross-validation.

Similarity. We quantify the representational similarity between X and Y by computing, for each split separately an *encoding score* with a Pearson correlation score $\mathrm{R} \in [-1, 1]^m$:

$$\mathrm{R} = corr(WX_{test}, y_{test})$$

We finally average these encoding scores across splits (and across sensors for MEG signals), leading to 1 encoding score for each subject in each brain location (fMRI) or each time sample (MEG).

Statistics. We compare each voxel's encoding score to a null distribution of 1000 encoding scores obtained with shuffled labels and apply FDR correction to account for the multiple voxels. We retain voxels with FDR-corrected $p < 0.05$. Only these significant voxels are displayed in the brain maps. For time-resolved MEG encoding, significance is assessed at each time point using a one-tailed, one-sample t-test, with FDR correction applied to account for multiple comparisons across time. Time points reaching $p < 0.05$ are indicated as significant with marks * in the figures.

2.2 MODELS

DINOv3. We used DINOv3 (small) as a model (Siméoni et al., 2025). This model is chosen as it is (1) state-of-the-art (2) self-supersived and (3) sufficiently small such that we can analyze the backpropagation gradients with our computational resources. A significant body of literature posits that the brain acts as an active inference engine, seeking to minimize variational free energy rather than merely maximizing external reinforcement. This framework suggests that the human brain relies on predictive coding to learn internal representations of the world through self-supervised error minimization, even in the absence of explicit reward signals (Friston and Kiebel, 2009). DINOv3 is selected as a modeling framework because its self-supervised objective may partially align with unsupervised learning principles at play in human brains. This model consists of a student network, optimized to match the representations of a teacher network. The teacher's weights are updated as an exponential moving average (EMA) of the student's weights, providing a slowly evolving, stable target. This setup allows the development of rich, invariant representations without labels. We train DINOv3-S from scratch to analyze the dynamics of Forward and Backward encoding scores across training.

Benchmark of eight models. We compare eight models spanning four architecture families: self-supervised ViTs (DINOv3-S (Siméoni et al., 2025), ViT-MAE (He et al., 2022)), a supervised ViT (ViT-L/16 (Dosovitskiy et al., 2020)), supervised CNNs (ConvNeXt-L (Liu et al., 2022), ResNet-50 (He et al., 2016)), and vision-language models (OWL-ViT (Minderer et al., 2022), CLIP (Radford et al., 2021), SigLIP2 (Tschannen et al., 2025)). We compute backpropagated gradients using each model's most natural loss: native reconstruction loss for ViT-MAE, native image-text similarity for VLMs, and CLIP-bridged caption similarity for classification models (ResNet-50, ConvNeXt-L, ViT-L/16). Gradients are collected from the residual block outputs at 10 evenly-spaced blocks throughout the vision encoder.

Forward activations and Backpropagated gradients. We extract the layer-wise activations of the model for each stimulus. Backpropagated gradients are computed with a standard implementation of

the algorithm between the teacher and student networks. Here, we leverage the image-wise DINOv3 losses, computed for each image seen, and backpropagated across the layers of the model.

Encoding of residual Backpropagation. In order to isolate the specific correlates of backpropagated gradients in the brain, we use a conservative method to remove correlated signals between Backpropagated gradients and Forward representations: we compute the Residual Backpropagation, defined as the encoding score of Forward activations and Backpropagated gradient, subtracted by the encoding score of Forward representation.

$$R_{\text{Residual Backprop}} = R(\text{Forward + Backprop}) - R(\text{Forward}) \quad (1)$$

Layer-wise extraction. We analyze 12 layers, indexed by their normalized depth $k \in [0, 1]$, where $k = 0$ and $k = 1$ correspond to the earliest and latest stages of the model, respectively. Specifically, we extract the activations from the projection layer (W3) of the MLP within each of the 12 transformer blocks. For each layer, we get the Forward $X_{forward} \in \mathbb{R}^{n \times d}$ and Backward signals $X_{backprop} \in \mathbb{R}^{n \times d'}$ and evaluate their linear mappings to the brain activity $Y \in \mathbb{R}^{n \times m}$ as the encoding score.

Random Projection and PCA. Backward signals have a high dimensionality of 4e6. To be able to store and analyze these signals within our computational resources, we perform projection of these signals onto a random matrix of dimension 800. Following the rationale of (Halko et al., 2011) this method approximates to a subspace PCA in high dimensionality. To mitigate the issue of heterogeneous vector sizes, we then transform these Forward $X_{forward} \in \mathbb{R}^{n \times d}$ and Backwards vectors $X_{backprop} \in \mathbb{R}^{n \times d'}$ with a Principal Component Analysis (n=300) with scikit-learn (Pedregosa et al., 2011) prior to encoding analyses. This preprocessing results in Forward $X_{forward} \in \mathbb{R}^{n \times 300}$ and Backward vectors $X_{backprop} \in \mathbb{R}^{n \times 300}$ of identical dimensions. To compute Residual Backpropagation, we compare the concatenation of $X_{forward} \in \mathbb{R}^{n \times 150}$ and $X_{backprop} \in \mathbb{R}^{n \times 150}$ to $X_{forward} \in \mathbb{R}^{n \times 300}$, allowing to compare features in the same dimensionality. We conduct a grid search across Projection dimensions and PCA dimensions to ensure the robustness of our results, Supp. Fig. 9.

2.3 DATASETS

Magnetoencephalography (MEG). We use the THINGS-MEG dataset (Hebart et al., 2023), which consists of MEG recordings from four healthy participants viewing 22,500 naturalistic images, representing a total of 1,800 object concepts. Images were presented during 1.5 s, while participants maintained fixation. To limit the impact of noise we apply a bandpass filter between 0.1 and 20 Hz, down-sample the signal at 30 Hz, time-lock the brain responses to individual images, and epoch the corresponding neural data between -0.5 s and +3 s relative to image onset using MNE-Python (Gramfort et al., 2013) and NeuroAI (King et al., 2026). Finally, we z-score MEG signals across images, for each MEG channel and each time point independently.

Functional Magnetic Resonance Imaging (fMRI). We leverage the Natural Scenes Dataset (Allen et al., 2022), a 7 tesla fMRI dataset which consists of recordings from eight subjects, each observing a total of 10 000 natural scenes during 4 seconds each, while performing a continuous recognition task. We encode fMRI beta responses time-locked to image onset, on the fsaverage surface. Across analyses, we study only voxels significantly encoded. **Best Encoded Regions of interest (ROIs).** We select the best-encoded ROIs, defined as those whose mean Forward encoding score falls within the top tercile (33%) of all mean ROI scores.

3 RESULTS.

Anatomical organization. We aim to identify whether forward and backward passes of a self-supervised vision model such as DINOv3 align with the anatomical organization of the human brain. Our results indicate that in the forward pass, early visual areas are best predicted by shallow layers, while associative cortex is best predicted by deeper layers (Fig. 2A), reflecting a bottom-up correspondence with cortical hierarchy. We find a significant positive correlation (Pearson $r = 0.35$, $p =$ 5e-3, Spearman $\rho = 0.27$, $p =$ 3e-2) between the distances of best-encoded ROIs from V1 and the layer that encode them best (Fig. 2C. These results replicate previous findings in the literature regarding a cohort of self-supervised and supervised models (Raugel et al., 2025; Wang et al., 2023).

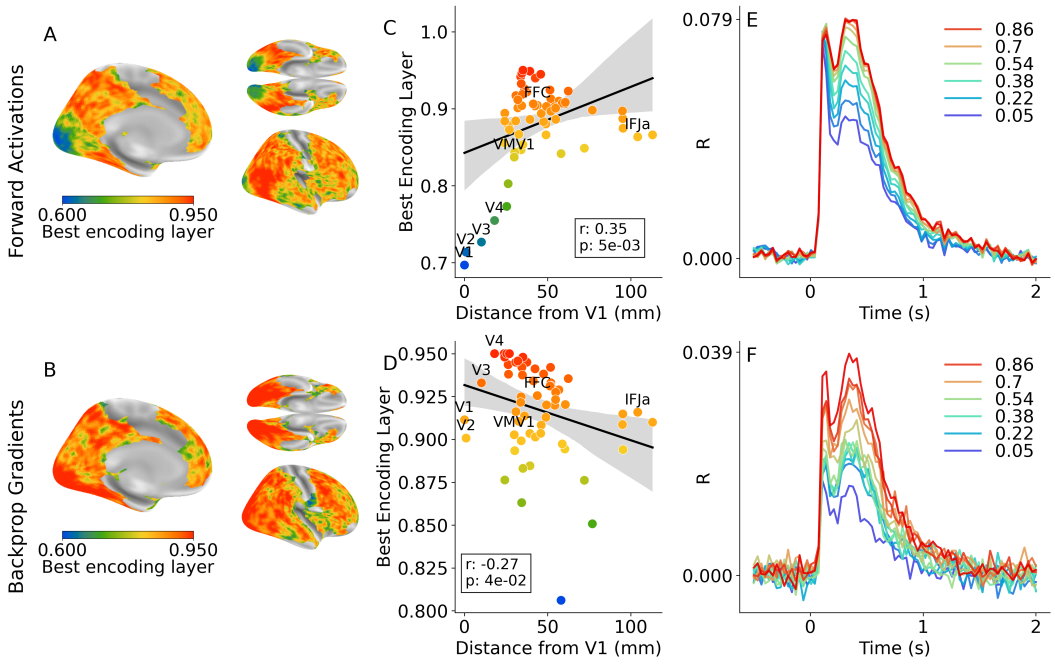


Figure 2: **Dissociation between forward and backward hierarchical alignment with the brain.** Panels A–B: fMRI cortical surface maps of the best encoding layer for Forward Activations and Backpropagated Gradients - approximated through PCA - at the end of training from scratch (at iteration 500k). Surface-smoothed (3 iterations, geodesic neighbor averaging, weight = 0.3). Panels C–D: Pearson correlation between ROI distance from V1 and best encoding layer, for the top third best encoded ROIs. Panels E–F: Time-resolved MEG encoding curves for each of the 12 layers.

Conversely, in the backward pass, early sensory areas are best aligned with deep gradients, whereas associative regions align with shallower gradients. (Fig. 2B). We find a negative spatial correlation (Pearson $r = -0.27$, $p =$ 4e-2, Spearman $\rho = -0.34$, $p =$ 8e-3) between the distances from V1 of the selected ROIs, and the layer that encode them best (Fig. 2F). This negative hierarchical trend indicates that gradient-based features do not follow the layer-to-cortical-area mapping observed in forward representations, and may even tend to reverse it.
Overall, the forward pass of DINOv3 aligns with the bottom-up anatomical organization of the human brain, whereas the backward pass does not align with the top-down organization of back-propagated gradients along layers of DINOv3.

Temporal organization. We aim to further compare the temporal organization of the activations in the human brain with the hierarchy of DINOv3, regarding feedforward and backpropagated signals. We first assess the temporal alignment of forward activations of DINOv3 with brain responses. The peaking times of encoding scores correspond to the order of their respective encoding layers: all layers peak around the same encoding score (R= 0.079), but shallower layers peak earlier, whereas deeper layers align later with brain responses (Fig. 2E).
Conversely, when considering backpropagated signals, peaks of alignment are not sequentially organized according to layer depth: all layers peak at the same time while a gradient in the peaking score is observed, with earlier layers peaking lower than later layers (Fig. 2F). This suggests that while feedforward activations in DINOv3 reflect a hierarchical temporal structure akin to that observed in the brain, backpropagated signals do not exhibit such a clear temporal hierarchy. Notably, back-propagated gradients across all layers peak simultaneously with the final layer of the forward pass. This lack of temporal delay contradicts biologically plausible implementations of backpropagation, which would predict backward signals to peak sequentially following the forward pass, Fig. 1.

Linear mapping of forward activations and backpropagated gradients is consistent across architectures. Backpropagated gradients from the self-supervised DINOv3-S model show signifi-

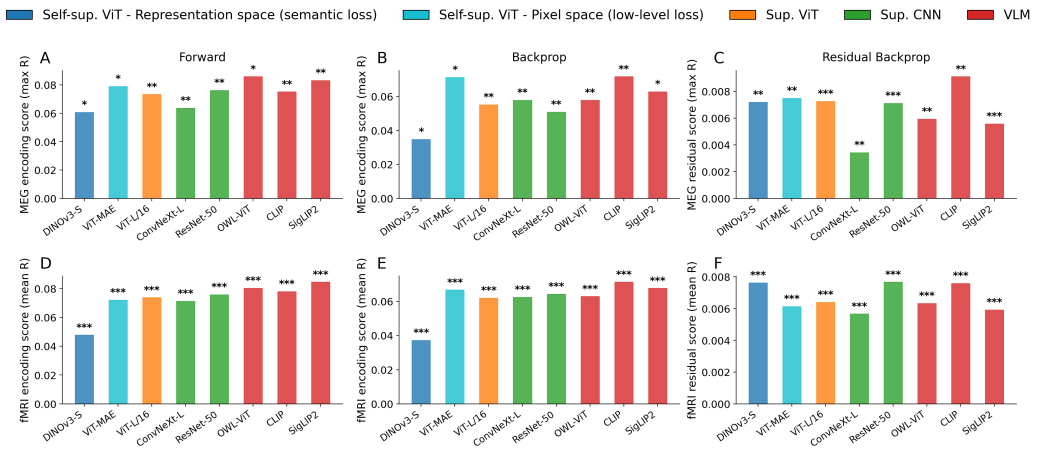


Figure 3: **Generalized neural alignment of forward and backward activations across diverse vision architectures.** Encoding scores of the last layer for eight vision models across Forward, Backprop, and Residual Backprop conditions. Top row: MEG encoding scores (maximum Pearson r across time of the mean-across-subjects temporal curve). Bottom row: fMRI encoding scores (mean Pearson r across significant voxels). Models are grouped by category: self-supervised ViTs trained with semantic loss (DINOv3-S) or pixel-level reconstruction loss (ViT-MAE), supervised ViT (ViT-L/16), supervised CNNs (ConvNeXt-L, ResNet-50), and vision-language models (OWL-ViT, CLIP, SigLIP2). Significancy: * $p < .05$, ** $p < .01$, *** $p < .001$.

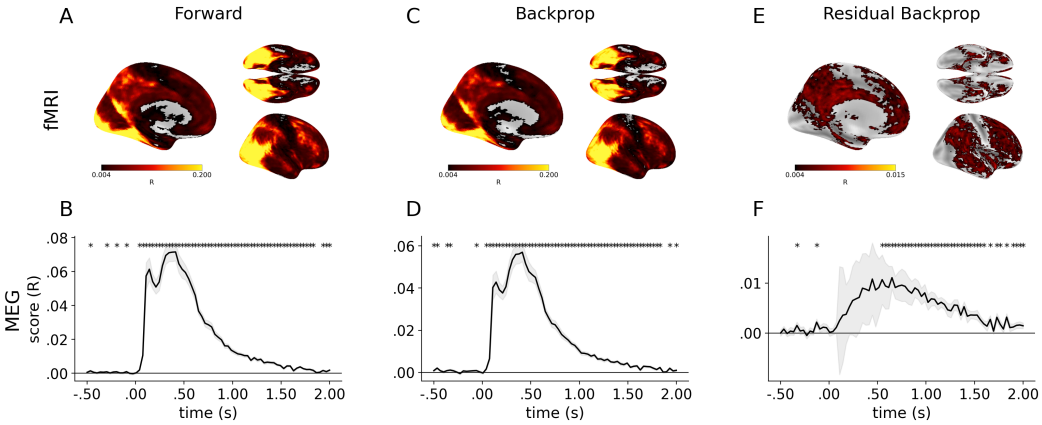


Figure 4: **Neural alignment over time and space with forward, backward and residual backward activations, averaged across models**. Top row: fMRI cortical surface maps showing the mean encoding score (Pearson r) across models for Forward (A), Backward (C), and Residual Backward (E) activations. Only significant voxels are plotted . Bottom row: MEG temporal encoding curves showing the mean score (black line) $\pm$ SEM (gray shading) across all subjects and models, for Forward (B), Backward (D), and Residual Backward (F) activations. Asterisks denote timepoints with significant encoding (FDR-corrected $p < 0.05$).

cant alignment with both hemodynamic (fMRI) and electrophysiological (MEG) brain responses, Fig. 2, 3. How does this mapping evolve across diverse architectures and learning objectives?

We compare eight models spanning four families of architectures and learning objectives: self-supervised ViTs (DINOv3-S, ViT-MAE), a supervised ViT (ViT-L/16), supervised CNNs (ConvNeXt-L, ResNet-50), and vision-language models (OWL-ViT, CLIP, SigLIP2). We investigate encoding scores of both forward and backward passes across processing time (MEG) and cortical space (fMRI).

First, results confirm that forward activations of these models can be mapped onto these brain record-

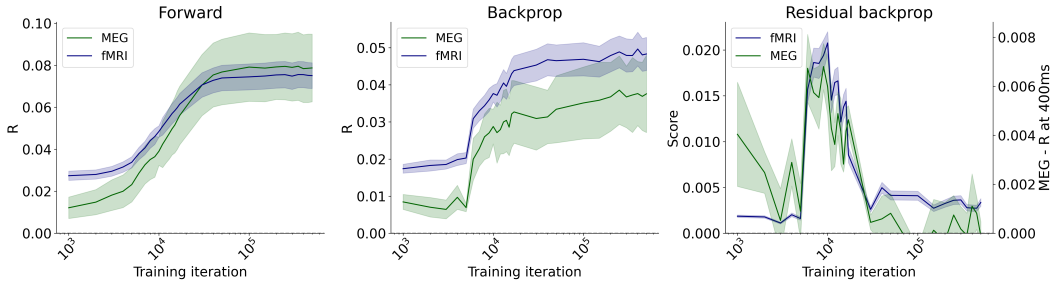


Figure 5: **Evolution of Forward, Backpropagation, and Residual Signals Across Model Training.** Results are shown for Forward (left), Backpropagation (middle), and Residual Backpropagation (right) encoding scores for the last layer. Both MEG (green) and fMRI (blue) scores exhibit synchronized trajectories across training iterations. Errorbars indicate the standard error across subjects for each modality.

ings (Fig. 3A, D, 4A, B). Forward activations significantly map onto brain responses in the visual system in fMRI, and shortly after image onset in MEG. This phenomenon was previously observed in the literature (Schrimpf et al., 2018; DiCarlo et al., 2012; Zhuang et al., 2021).
We additionally find that backpropagated gradients of each of these models also significantly map onto the human brain recordings, Fig. 3B, E. Similar regions of the visual system and processing latencies are engaged by both Forward and Backprop activity, indicating that Backprop specifically encodes the synaptic updating occurring in these visual regions, Fig. 4C, D.
Finally, we compute the Residual Backpropagation, a conservative measure which isolates gradient-specific neural variance beyond Forward activations. Residual Backpropagation scores are obtained from backpropagated features after removing the variance explained by Forward activations. Residual Backpropagation still significantly predicts brain activity, demonstrating that backpropagated gradients capture variance in neural responses that is not accounted for by Forward activations alone, Fig. 3C, F.
Interestingly, Residual Backpropagation encodes most accurately representations of the late visual cortex, Fig. 4E. This observation is corroborated by MEG results, where residual backpropagation demonstrates its highest encoding performance during later latencies, between 0.5s and 1,7s after image onset, Fig. 4F. Taken together, these results indicate that residual backpropagation primarily align with higher-level representations in the late visual cortex.
Overall, these results show that backpropagated gradients significantly map onto brain activity during passive viewing of images and explain neural responses beyond what is captured by Forward activations. This finding suggests that the brain might implement learning mechanisms that share similarities with backpropagation, such as predictive coding (Lillicrap et al., 2020; Bastos et al., 2012).// Additionally, we reproduce the temporal and spatial gradients observed in Fig. 2 for Forward and Backward activations, combining these eight models, Supp. Fig. 8.

Training Dynamics of Forward Activations, Backpropagated Gradients and Residual Backpropagation. We wonder how Forward and Backpropagated signals align with the human brain across model training. To study this question, we train from scratch DINOv3-S on naturalistic data. We chose DINOv3 under the assumption that its SSL objective may be the closest approximate – among the eight studied models – of the brain's capacity for representation learning through passive, unsupervised environmental exposure. Encoding scores from forward and backpropagated signals increase logarithmically throughout training from scratch, exhibiting consistent trajectories across both MEG and fMRI modalities, Fig. 5. Both Forward and Backpropagation scores expand across training towards the higher-level regions of the cortex, and towards later latencies, Fig. 6, 7.
In contrast, residual backpropagation scores follow a distinct non-monotonic trajectory: they initially increase, reach a distinct peak at 10k iterations, and subsequently decrease. This pattern is consistent across both neural recording methods fMRI and MEG, demonstrating a synchronized evolution across both temporal and spatial scales. At peak time (10k iterations), Residual Backpropagation significantly encodes late-latency neural activations, and associative brain regions in the late visual cortex, involved in higher-level processing – ventromedial and lateral-occipitotemporal cortex. This transient peak suggests the emergence a learning signal that intensifies across training as

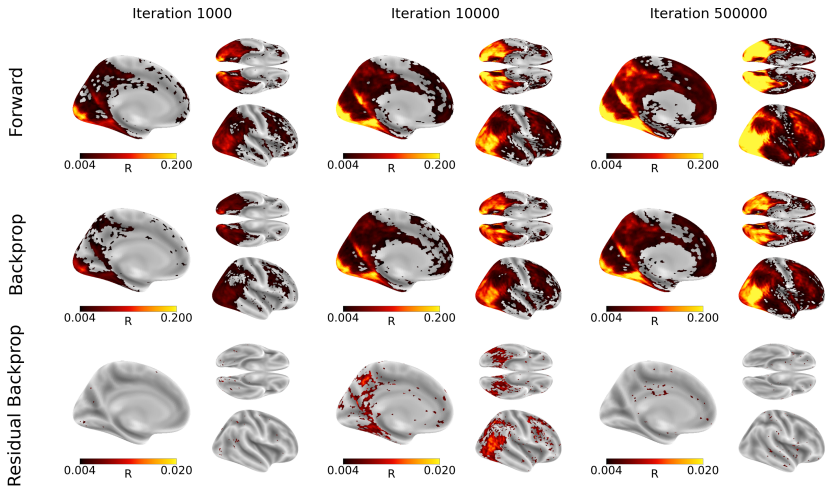


Figure 6: **Spatial Distribution of Forward, Backpropagation, and Residual Backpropagation across training.** FMRI encoding scores for Forward Activations (up), Backpropagated Gradients (middle), and Residual Backpropagated Gradients (Bottom) at three training stages for the deeper layer. Only significant voxels ($p < 0.05$) are shown.

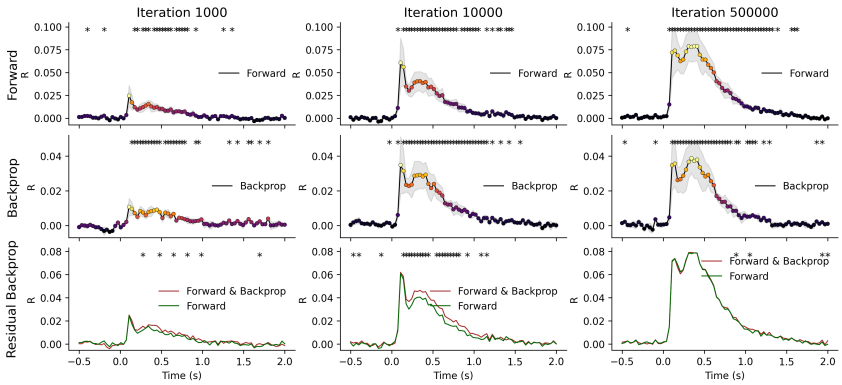


Figure 7: **Temporal Distribution of Forward, Backpropagation, and Residual Signals across training.** MEG encoding scores for Forward Activations (up), Backpropagated Gradients (middle), and Residual Backpropagated Gradients (Bottom) at three training stages for the deeper layer. Shaded areas represent the standard error across subjects. Asterisks denote timepoints with significant encoding ($p < 0.05$).

gradients align with meaningful updates, then subsides as the DINOv3 teacher and student networks converge.

DISCUSSION

Summary of findings. In this study, we investigated the alignment between DINOv3 (Siméoni et al., 2025), a SOTA self-supervised vision transformer, and the human brain response to natural images. We found that backpropagated gradients from DINOv3 can be linearly mapped onto fMRI and MEG responses to natural images, indicating that gradient-related signals are reflected in brain activity. These results were replicated across 8 diverse vision models.

Importantly, even after removing the contribution of forward activations, backpropagated gradients still explained a significant portion of neural responses. This indicates that gradients capture additional variance in brain activity beyond forward representations, specifically for later latencies and in higher-level ROIs – e.g. multisensory integration, memory, semantic processing. This alignment

may reflect intensive synaptic updating regarding high-level representation.
However, the temporal and spatial organization of this mapping does not match predictions from a biologically plausible implementation of backpropagation. First, backward encoding does not peak later than forward encoding. Second, contrary to what can be observed with the forward activations, the expected top-down hierarchy of backpropagated gradients does not match the anatomical and temporal organization of brain responses.

Relation to biological backpropagation theories. A substantial body of prior work has examined how backpropagation, or functionally equivalent learning rules, could in principle be implemented in biological neural networks. Much of this effort has focused on resolving the weight transport problem – the requirement that feedback connections mirror feedforward weights – which is not supported by cortical circuitry. Proposed solutions, including random feedback alignment and segregated feedforward and feedback pathways, demonstrate that precise synaptic symmetry is not strictly necessary for gradient-based learning (Lillicrap et al., 2020). Importantly, these variants preserve the core computational structure of backpropagation: error signals are computed at higher levels and propagated downward to guide synaptic updates at earlier stages. Our findings do not contradict the theoretical feasibility of such learning mechanisms. Rather, they provide empirical constraints, suggesting that if a gradient-based learning operates in the brain, it does not manifest with the spatial and temporal signatures predicted by backpropagation.

Implications for learning mechanisms in the brain. The observed dissociation between representational similarity and learning dynamics suggests that shared representations between deep networks and the brain do not necessarily imply shared learning algorithms. While forward activations in deep models reliably mirror cortical hierarchies, the organization of gradient-related signals does not follow the canonical bottom-up / top-down sequence implied by backpropagation. This raises the possibility that the brain relies on alternative learning mechanisms that give rise to similar representational structures without explicitly computing or propagating error gradients. Candidate frameworks include local Hebbian or contrastive learning rules (Zenke et al., 2017), predictive coding architectures (Bastos et al., 2012; Oord et al., 2018), and other biologically grounded credit-assignment strategies. Discriminating among these possibilities will require models that make explicit, testable predictions about both neural activity and learning dynamics.

Modeling limitations. While representational mapping was evaluated on a broad set of eight models, the in-depth training dynamics analysis was restricted to a single SSL model, DINOv3-S, chosen for its strong performance and well-established correspondence with visual cortical representations. Focusing on a SSL model ensures that the learning mechanism could, in principle, be biologically plausible, as it does not require additional information, such as the image labels. However, learning mechanisms of DINOv3 such as the use of EMA and the batch-based nature of backpropagation may not directly map onto human learning strategies. Additionally, unlike the brain, DINOv3 operates on static images and thus lacks explicit temporal dynamics or recurrent processing, limiting its ability to capturing learning signals that unfold over space rather than over time. Future research should aim to refine these models to better mimic biological learning processes and further elucidate the computational principles underlying brain function. Future work should also extend this approach to a broader range of architectures, including recurrent or predictive models, and video-based systems that use backpropagation with a different type of prediction objective (e.g. predict the next frame).

Measurement limitations. Our conclusions are also constrained by the spatial and temporal resolution of non-invasive human neuroimaging. While fMRI and MEG provide complementary coverage of spatial and temporal dynamics, they cannot directly resolve laminar-specific activity or synapse-level learning signals. Invasive recordings or animal models with controlled learning paradigms may be necessary to more directly probe the neural implementation of credit assignment.

Impact statement. The human brain remains extraordinarily more data – and energy – efficient than recent deep networks. Understanding its learning mechanisms is thus a key goal for both neuroscience and AI. The divergence between the human brain and the backward parts of a SOTA self-supervised model (Siméoni et al., 2025) suggests that there exists a different – potentially more efficient – strategy to update a large number of synaptic weights. Understanding this biological algorithm could inspire novel, more efficient, robust, and generalizable approaches to machine learning, and - perhaps more importantly - this approach could illuminate the biological principles of intelligence.

REFERENCES

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL http://dx.doi.org/10.1038/s41593-021-00962-x.

Andre M. Bastos, W. Martin Usrey, Rick A. Adams, George R. Mangun, Pascal Fries, and Karl J. Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, November 2012. ISSN 0896-6273. doi: 10.1016/j.neuron.2012.10.038. URL http://dx.doi.org/10.1016/j.neuron.2012.10.038.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–462, 2014.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL https://arxiv.org/abs/2010.11929.

Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221, May 2009. ISSN 1471-2970. doi: 10.1098/rstb.2008.0300. URL http://dx.doi.org/10.1098/rstb.2008.0300.

Shohei Furutachi and Sonja B. Hofer. Rethinking predictive processing. *Annual Review of Neuroscience*, 49(1):471–494, 2026. ISSN 1545-4126. doi: 10.1146/annurev-neuro-102124-031410. URL http://dx.doi.org/10.1146/annurev-neuro-102124-031410.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.

U. Guclu and M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, July 2015. ISSN 1529-2401. doi: 10.1523/jneurosci.5023-14.2015. URL http://dx.doi.org/10.1523/JNEUROSCI.5023-14.2015.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, January 2011. ISSN 1095-7200. doi: 10.1137/090771806. URL http://dx.doi.org/10.1137/090771806.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 770–778. IEEE, 2016. doi: 10.1109/cvpr.2016.90. URL http://dx.doi.org/10.1109/CVPR.2016.90.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 15979–15988. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01553. URL http://dx.doi.org/10.1109/CVPR52688.2022.01553.

Martin N. Hebart, Oliver Contier, Lina Teichmann, Adam H. Rockter, Charles Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I. Baker. "things-meg", 2023.

Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P De Lange. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119, 2022.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11):e1003915, November 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003915. URL http://dx.doi.org/10.1371/journal.pcbi.1003915.

J-R. King, C. Bel, L. Evanson, J. Gadonneix, S. Houhamdi, J. Lévy, J. Raugel, A. Santos Revilla, M. Zhang, J. Bonnaire, C. Caucheteux, A. Défossez, T. Desbordes, P. Diego-Simón, S. Khanna, J. Millet, P. Orhan, S. Panchavati, A. Ratouchniak, A. Thual, T. Brooks, K. Begany, Y. Benchetrit, M. Careil, H. Banville, S. d'Ascoli, S. Dahan, and J. Rapin. Neuralset: A high-performing python package for neuro-ai. 2026. URL https://kingjr.github.io/files/neuralset.pdf. Preprint; URL will be updated when the paper lands on arXiv.

Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey E Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21:335–346, 2020.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 11966–11976. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01167. URL http://dx.doi.org/10.1109/CVPR52688.2022.01167.

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning, 2023. URL https://arxiv.org/abs/2206.01685.

Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. Predictive coding approximates backprop along arbitrary computation graphs. *Neural Computation*, 34(6):1329–1368, May 2022. ISSN 1530-888X. doi: 10.1162/neco_a_01497. URL http://dx.doi.org/10.1162/neco_a_01497.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. *Simple Open-Vocabulary Object Detection*, page 728–755. Springer Nature Switzerland, 2022. ISBN 9783031200809. doi: 10.1007/978-3-031-20080-9_42. URL http://dx.doi.org/10.1007/978-3-031-20080-9_42.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410, 2011.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, January 1999. ISSN 1546-1726. doi: 10.1038/4580. URL http://dx.doi.org/10.1038/4580.

Joséphine Raugel, Marc Szafraniec, Huy V. Vo, Camille Couprie, Patrick Labatut, Piotr Bojanowski, Valentin Wyart, and Jean-Rémi King. Disentangling the factors of convergence between brains and computer vision models, 2025. URL https://arxiv.org/abs/2508.18226.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. ISSN 1476-4687. doi: 10.1038/323533a0. URL http://dx.doi.org/10.1038/323533a0.

Yuka Sasaki, Jose E. Nanez, and Takeo Watanabe. Advances in visual perceptual learning and plasticity. *Nature Reviews Neuroscience*, 11(1):53–60, December 2009. ISSN 1471-0048. doi: 10.1038/nrn2737. URL http://dx.doi.org/10.1038/nrn2737.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, November 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00753-y. URL http://dx.doi.org/10.1038/s42256-023-00753-y.

James C. R. Whittington and Rafal Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Computation*, 29(5):1229–1262, May 2017. ISSN 1530-888X. doi: 10.1162/neco_a_00949. URL http://dx.doi.org/10.1162/NECO_a_00949.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, May 2014. ISSN 1091-6490. doi: 10.1073/pnas.1403112111. URL http://dx.doi.org/10.1073/pnas.1403112111.

Friedemann Zenke, Wulfram Gerstner, and Surya Ganguli. The temporal paradox of hebbian learning and homeostatic plasticity. *Current opinion in neurobiology*, 43:166–176, 2017.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), January 2021. ISSN 1091-6490. doi: 10.1073/pnas.2014196118. URL http://dx.doi.org/10.1073/pnas.2014196118.

SUPPLEMENTARY MATERIALS

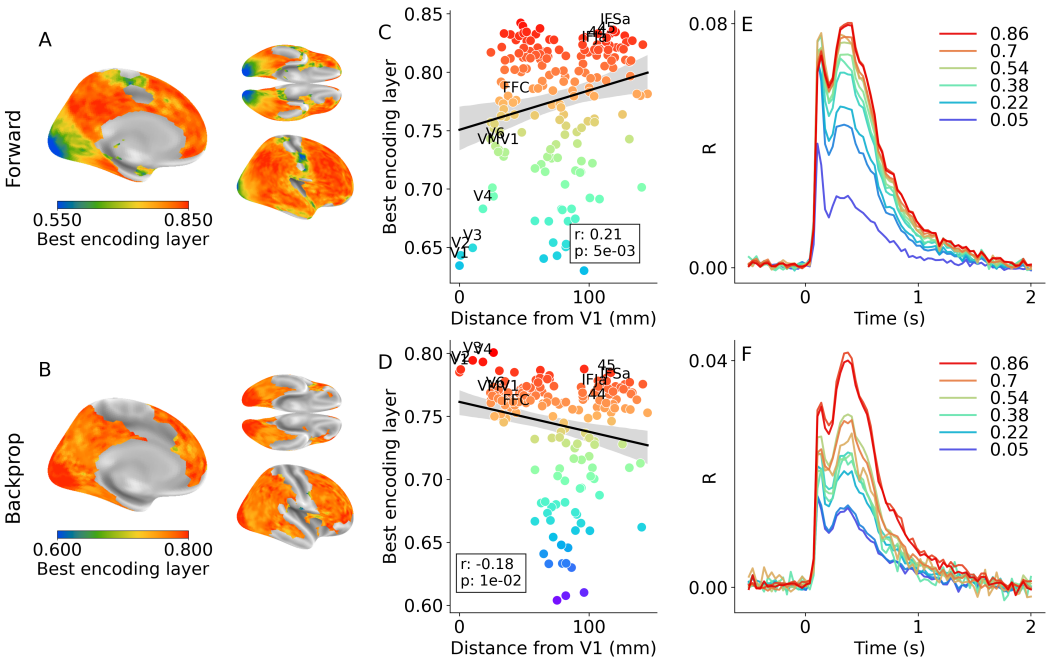


Figure 8: DINOv3-S encoding across all layers, combined across a combination of eight vision models spanning diverse architectures. Panels A–B: fMRI cortical surface maps of the best encoding layer (averaged across the three fMRI datasets) for Forward and Backprop representations. Surface-smoothed (3 iterations, geodesic neighbor averaging, weight = 0.3). Only significant voxels plotted (FDR-corrected $p < 0.05$ among the 8 model scores). Panels C–D: Pearson correlation between each ROI's distance from V1 and its best encoding layer. Panels E–F: MEG/EEG temporal encoding curves for each layer (averaged across subjects from all three MEG/EEG datasets), for Forward and Backprop.

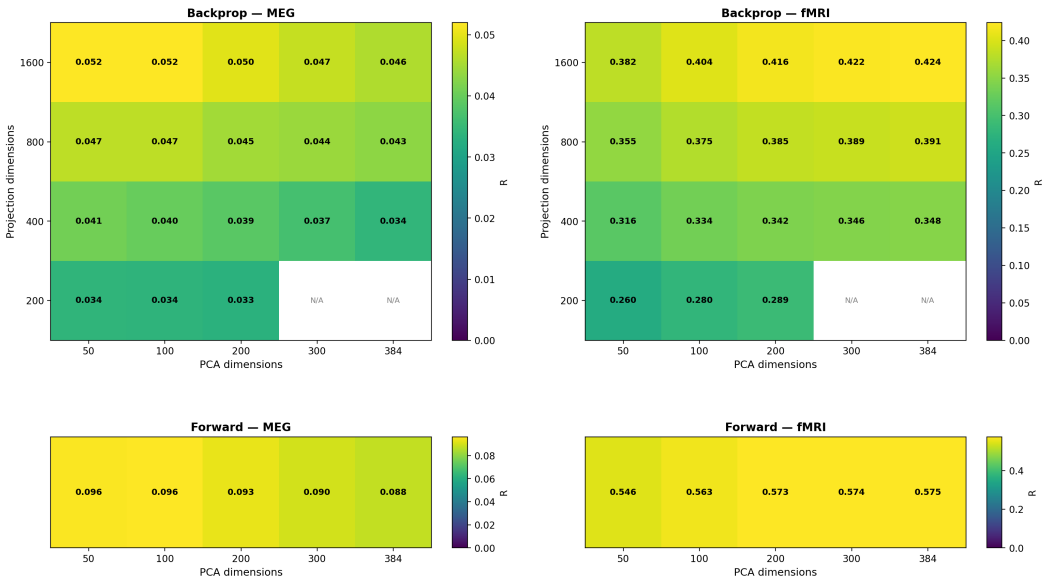


Figure 9: **Robustness to PCA dimensionality and projection dimension.** (Top row) Backprop encoding score as a function of PCA dimensions (x-axis) and random projection dimensions (y-axis) for MEG (left) and fMRI (right). (Bottom row) Forward encoding score vs. PCA dimensions (no random projection applied).

AI USE STATEMENT

Generative AI tools were used to assist with language editing and code development, all content verified by the authors. All scientific ideas, analyses, results, interpretations, and conclusions were developed and verified by the authors, who take full responsibility for the content of the paper.

ETHICS STATEMENT

This work analyzes previously published, publicly available human neuroimaging datasets ((Allen et al., 2022) and (Hebart et al., 2023)) and introduces no new data collection or intervention involving human participants. We follow the ethical requirements associated with these datasets and the ICLR Code of Ethics. We are not aware of additional ethical risks specific to this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the datasets, models, preprocessing, feature extraction, encoding analyses, statistical procedures, and hyperparameters in the Methods. Code necessary to reproduce the analyses is open-source: MNE-Python (Gramfort et al., 2013) and NeuroAI (King et al., 2026).

REFERENCES

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL http://dx.doi.org/10.1038/s41593-021-00962-x.

Andre M. Bastos, W. Martin Usrey, Rick A. Adams, George R. Mangun, Pascal Fries, and Karl J. Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, November 2012. ISSN 0896-6273. doi: 10.1016/j.neuron.2012.10.038. URL http://dx.doi.org/10.1016/j.neuron.2012.10.038.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–462, 2014.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL https://arxiv.org/abs/2010.11929.

Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221, May 2009. ISSN 1471-2970. doi: 10.1098/rstb.2008.0300. URL http://dx.doi.org/10.1098/rstb.2008.0300.

Shohei Furutachi and Sonja B. Hofer. Rethinking predictive processing. *Annual Review of Neuroscience*, 49(1):471–494, 2026. ISSN 1545-4126. doi: 10.1146/annurev-neuro-102124-031410. URL http://dx.doi.org/10.1146/annurev-neuro-102124-031410.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.

U. Guclu and M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, July 2015. ISSN 1529-2401. doi: 10.1523/jneurosci.5023-14.2015. URL http://dx.doi.org/10.1523/JNEUROSCI.5023-14.2015.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, January 2011. ISSN 1095-7200. doi: 10.1137/090771806. URL http://dx.doi.org/10.1137/090771806.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 770–778. IEEE, 2016. doi: 10.1109/cvpr.2016.90. URL http://dx.doi.org/10.1109/CVPR.2016.90.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 15979–15988. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01553. URL http://dx.doi.org/10.1109/CVPR52688.2022.01553.

Martin N. Hebart, Oliver Contier, Lina Teichmann, Adam H. Rockter, Charles Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I. Baker. "things-meg", 2023.

Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P De Lange. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119, 2022.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11):e1003915, November 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003915. URL http://dx.doi.org/10.1371/journal.pcbi.1003915.

J-R. King, C. Bel, L. Evanson, J. Gadonneix, S. Houhamdi, J. Lévy, J. Raugel, A. Santos Revilla, M. Zhang, J. Bonnaire, C. Caucheteux, A. Défossez, T. Desbordes, P. Diego-Simón, S. Khanna, J. Millet, P. Orhan, S. Panchavati, A. Ratouchniak, A. Thual, T. Brooks, K. Begany, Y. Benchetrit, M. Careil, H. Banville, S. d'Ascoli, S. Dahan, and J. Rapin. Neuralset: A high-performing python package for neuro-ai. 2026. URL https://kingjr.github.io/files/neuralset.pdf. Preprint; URL will be updated when the paper lands on arXiv.

Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey E Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21:335–346, 2020.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 11966–11976. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01167. URL http://dx.doi.org/10.1109/CVPR52688.2022.01167.

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning, 2023. URL https://arxiv.org/abs/2206.01685.

Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. Predictive coding approximates backprop along arbitrary computation graphs. *Neural Computation*, 34(6):1329–1368, May 2022. ISSN 1530-888X. doi: 10.1162/neco_a_01497. URL http://dx.doi.org/10.1162/neco_a_01497.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. *Simple Open-Vocabulary Object Detection*, page 728–755. Springer Nature Switzerland, 2022. ISBN 9783031200809. doi: 10.1007/978-3-031-20080-9_42. URL http://dx.doi.org/10.1007/978-3-031-20080-9_42.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410, 2011.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, January 1999. ISSN 1546-1726. doi: 10.1038/4580. URL http://dx.doi.org/10.1038/4580.

Joséphine Raugel, Marc Szafraniec, Huy V. Vo, Camille Couprie, Patrick Labatut, Piotr Bojanowski, Valentin Wyart, and Jean-Rémi King. Disentangling the factors of convergence between brains and computer vision models, 2025. URL https://arxiv.org/abs/2508.18226.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. ISSN 1476-4687. doi: 10.1038/323533a0. URL http://dx.doi.org/10.1038/323533a0.

Yuka Sasaki, Jose E. Nanez, and Takeo Watanabe. Advances in visual perceptual learning and plasticity. *Nature Reviews Neuroscience*, 11(1):53–60, December 2009. ISSN 1471-0048. doi: 10.1038/nrn2737. URL http://dx.doi.org/10.1038/nrn2737.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, November 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00753-y. URL http://dx.doi.org/10.1038/s42256-023-00753-y.

James C. R. Whittington and Rafal Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Computation*, 29(5):1229–1262, May 2017. ISSN 1530-888X. doi: 10.1162/neco_a_00949. URL http://dx.doi.org/10.1162/NECO_a_00949.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, May 2014. ISSN 1091-6490. doi: 10.1073/pnas.1403112111. URL http://dx.doi.org/10.1073/pnas.1403112111.

Friedemann Zenke, Wulfram Gerstner, and Surya Ganguli. The temporal paradox of hebbian learning and homeostatic plasticity. *Current opinion in neurobiology*, 43:166–176, 2017.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), January 2021. ISSN 1091-6490. doi: 10.1073/pnas.2014196118. URL http://dx.doi.org/10.1073/pnas.2014196118.