

1 Directly interfacing brain and deep networks 2 exposes non-hierarchical visual processing

3 Nicholas J. Sexton^{*†} Bradley C. Love^{†‡}

4 March 2021

5 **Abstract**

6 One reason the mammalian visual system is viewed as hierarchical,
7 such that successive stages of processing contain ever higher-level infor-
8 mation, is because of functional correspondences with deep convolutional
9 neural networks (DCNNs). However, these correspondences between brain
10 and model activity involve shared, not task-relevant, variance. We pro-
11 pose a stricter test of correspondence: If a DCNN layer corresponds to
12 a brain region, then replacing model activity with brain activity should
13 successfully drive the DCNN's object recognition decision. Using this ap-
14 proach on three datasets, we found all regions along the ventral visual
15 stream best corresponded with later model layers, indicating all stages
16 of processing contained higher-level information about object category.
17 Time course analyses suggest long-range recurrent connections transmit
18 object class information from late to early visual areas.

19 **1 Introduction**

20 Despite some shortcomings(1), deep convolutional neural networks (DCNNs)
21 have emerged as the best candidate models for the mammalian visual system.
22 These models take photographic stimuli as input and, after traversing multiple
23 layers consisting of millions of connection weights, output a class or category
24 label. Weights are trained on large datasets consisting of natural images and
25 corresponding labels.

26 The deep learning revolution in neuroscience began when layers of DCNNs
27 were related to regions along the ventral visual stream in an early-to-early and
28 late-to-late pattern of correspondence between brain regions and model layers
29 (2-4) (fig 1A). This correspondence supported the view that the ventral stream
30 is a hierarchy in which ever more complex features and higher-level information

*Correspondence to: n.sexton@ucl.ac.uk

†Department of Experimental Psychology, University College London, London, United Kingdom.

‡The Alan Turing Institute, London, United Kingdom

31 are encoded as one moves from early visual areas like V1 or V4 to inferotemporal
32 (IT) cortex (5).

33 However, these correspondences between brain and model activity were based
34 on total shared variance as opposed to task-relevant variance (fig 1B). Much of
35 cortex-wide neural variance does not relate to the task of interest(6) and may
36 co-vary with but not drive behaviour. Correspondences established by correla-
37 tion alone do not necessitate that model layers and brain regions play the same
38 functional role in the overall computation.

39 We propose a stronger test for evaluating how brain-like a model is. If, as
40 is frequently claimed(2-4), a specific layer in a DCNN corresponds to a brain
41 region, then it should be possible to substitute the activations on that layer
42 with the corresponding brain activity and drive the DCNN to an appropriate
43 output (cf. (7, 8), fig 1C). For example, if we take V4 activity from a monkey
44 viewing an image of a car and interface that brain activity with an intermediate
45 DCNN layer hypothesised to correspond to V4, then the DCNN should respond
46 “car” absent any image input. How well the DCNN performs when directly
47 interfaced (through a simple linear mapping, see SI 6.5) with the brain provides
48 a strong test of how well the interfaced brain region corresponds to that layer
49 of the DCNN.

50 2 Driving model response with brain activity

51 We interfaced a pretrained DCNN(9) with data from two human brain imaging
52 studies(10, 11) and a Macaque monkey study(12). All three studies involved
53 viewing complex images. For a chosen model layer and brain region, we cal-
54 culated a linear mapping from brain to model activity by presenting the same
55 images to the model for which we had neural recordings (fig 1C). This simple
56 linear mapping is a translation between brain and model activity. We evaluated
57 the quality of this translation by considering held-out images and brain data
58 that were not used in calculating the linear mapping (see SI 6.4).

59 Strikingly, for the two fMRI studies (figs 2A, 2B), the DCNN was most ac-
60 curate at classifying novel images when brain activity across regions (both early
61 and late along the ventral stream) was interfaced with later model layers. In
62 contrast to previous analyses that focused on total variance, we did not find the
63 early-to-early and late-to-late pattern of correspondence. Even primary visual
64 cortex, V1, best drove the DCNN when interfaced with an advanced layer. For
65 comparison, classifiers commonly used to decode information from fMRI data
66 through multivariate pattern analysis (MVPA) were at chance levels (fig 6),
67 which highlights the useful constraints captured in the pretrained DCNN. After
68 training on a million naturalistic images, the DCNN developed representations
69 that paralleled those of the ventral stream, which made decoding object class
70 possible by way of a linear mapping from brain activity to an advanced DCNN
71 layer. The interpretation is that all brain regions contain advanced object recog-
72 nition information, which conflicts with strict hierarchical views of the ventral
73 visual stream.

74 To rule out any alternative explanation based on the indirect nature of fMRI
75 recordings, we considered a third study consisting of direct multi-unit recording
76 of spiking neurons implanted in the ventral visual stream of Macaque mon-
77 keys(12). These monkeys were shown images that did not readily align with
78 the pretrained DCNN’s class labels, so we evaluated neural translation perfor-
79 mance by comparing the outputs of the DCNN when its input was a study
80 image vs. when a DCNN layer was driven by brain data elicited by the same
81 image. For the distance measure, KL divergence, lower values imply a better
82 translation between brain and model activity. As in the fMRI studies, both
83 relatively early regions (i.e., V4) and late regions (i.e., IT) best translated to
84 later DCNN layers (fig 2C).

85 Across three diverse studies, we found a remarkably consistent pattern that
86 strongly diverged from previous analyses — both early and late regions along the
87 ventral visual stream best corresponded (i.e., translated) to late model layers. It
88 is not that previous analyses were poorly conducted (see SI fig 5 for a successful
89 reanalysis of data(12) finding the early-to-early and late-to-late canonical pat-
90 tern). Rather, our novel analyses focused on task-relevant analysis, i.e., variance
91 that can drive behaviour, provided a different view of the system than standard
92 analyses focused on shared variance. Integrating these two views suggests a
93 non-hierarchical account of object recognition marked by long-range recurrence
94 transmitting higher-level information to the earliest visual areas.

95 **3 Long-range recurrence as opposed to strict hi-** 96 **erarchy**

97 One way to reconcile the existing literature based on shared variance with our
98 analyses based on task-relevant variance is to propose that long-range connec-
99 tions from IT transmit higher-level information to early visual areas. Even
100 if most variance in lower-level visual areas is attributable to stimulus-driven,
101 bottom-up activity, the majority of task-relevant information could be attributable
102 to signals originating from IT (fig 3).

103 This view predicts specific patterns of Granger causality between early and
104 late areas along the ventral visual stream. Do past values of one time series
105 predict future values of the other? In terms of total spiking activity, lower-level
106 areas should first cause activity in higher-level areas during the initial feed-
107 forward pass in which stimulus-driven activity propagates along the ventral
108 visual stream. Later in processing, the causality should become reciprocal as
109 top-down connections from IT affect firing rates in lower-level areas, such as V4
110 (fig 3, bottom row). In contrast, Granger causality for task-relevant information
111 should first be established from IT to V4 (i.e., the top-down signal) and only
112 later in processing should recurrent activity lead to causality from V4 to IT
113 (fig 3, top row). In this fashion, all areas are effectively “late” after long-range
114 recurrent connections transmit information from IT to early visual areas along
115 the ventral stream though most variance for these areas would be dominated by

116 lower-level (bottom-up) stimulus information.

117 We tested these predictions using the monkey multi-unit spiking data(12)
118 that has the temporal resolution to support the analyses. Images were presented
119 one after the other, each visible for 100ms, with a 100ms period between stimuli.
120 Figure 4A shows the mean firing rates (10 ms bins) with activity in V4 increasing
121 shortly before IT, consistent with stimulus-related activity first occurring in V4.
122 Figure 4B revisits our previous analyses (fig 2C) but with spike counts binned
123 into 10ms intervals rather than aggregated over the entire trial. Even with only
124 10ms of recordings, neural translation from V4 and IT to an advanced DCNN
125 network layer minimises KL divergence between model outputs arising from
126 image input vs. when driven by brain activity.

127 Turning to the key Granger causality analyses, we evaluated whether early
128 ventral stream regions become more like late-ventral stream regions over time
129 due to recurrence (fig. 3). As processing unfolded, we found mutual causality
130 between lower-level (V4) and higher-level (IT) areas for analyses conducted over
131 spike counts (fig 4C) and for analyses on the KL divergence time series that
132 assessed the ability of brain regions to drive DCNN response (fig 4D).

133 Critically, the specific predictions of the long-range recurrence hypothesis
134 were supported with V4 first driving IT ($V4 \rightarrow IT$) for the analysis of spike
135 counts but IT first driving V4 ($V4 \leftarrow IT$) for the task-relevant information
136 analysis using the KL divergence time series (see SI for details). These results
137 are consistent with stimulus-driven bottom-up activity proceeding from V4 to
138 IT on an initial feed forward pass through the ventral stream with actionable
139 information about object recognition first arising in IT. Then, recurrent con-
140 nections from IT to V4 make task-relevant information available to V4. As
141 this loop is completed and cycles, both areas mutually influence one another
142 with the impact of bottom-up stimulus information maintained throughout the
143 process.

144 4 Discussion

145 Computational models can help infer the function of brain regions by linking
146 model and brain activity. Multilayer models, such as DCNNs, are particularly
147 promising in this regard because their layers can be systematically mapped
148 to brain regions. Indeed, the deep learning revolution in neuroscience began
149 with analyses suggesting an early-to-early, late-to-late pattern of correspondence
150 between DCNN layers and brain regions along the ventral visual stream during
151 object recognition tasks(2-4).

152 However, as we have argued, correspondences based on total shared variance
153 should be treated with caution. To complement these approaches, we presented
154 a test focused on task-relevant variance that directly interfaced neural recordings
155 with a DCNN model. If a brain region corresponds functionally to a model layer,
156 then brain activity substituted for model activity at that layer should drive the
157 model to the same output as when an image stimulus is presented. Of course,
158 models and brains speak different languages, so a translation between brain and

159 model activity must first be learned, which in our case was accomplished by
160 a linear transformation. Once the translation function is learned, novel brain
161 data and images can be used to evaluate possible brain-model correspondences.

162 Our approach, which focuses on task relevant variance within the overall
163 computation, as opposed to local shared variance (fig 1), uncovered a pattern
164 of correspondences that dramatically differed from the existing literature. We
165 found that all brain regions, from the earliest to the latest of visual areas along
166 the ventral stream, best corresponded to later model layers. These results indi-
167 cate that neural recordings in all regions contain higher-level information about
168 object category even when most variance in a region is attributable to lower-level
169 stimulus properties (fig 3).

170 To resolve this discrepancy between our analyses focused on task-relevant
171 variance and those based on shared variance, we evaluated the hypothesis that
172 long-range recurrence between higher-level brain regions, such as IT, influenced
173 activity in lower-level areas like V4. Analysing both firing rates of cells and
174 information-level analyses using our brain-model interface approach, we found
175 evidence that recurrent activity renders all areas functionally “late” as process-
176 ing unfolds, even when total variance in some early visual regions is largely
177 driven by bottom-up stimulus information. In this way, we integrate previ-
178 ous findings with our own and highlight how our method can be used to test
179 hypotheses about information flow in the brain.

180 Our approach, which considers task-relevant variance, may help resolve con-
181 flicting interpretations on the function of brain regions. For example, the
182 fusiform face area (FFA) responds selectively for faces, but its wider functional
183 role in object recognition has been the subject of extensive debate(13). Here,
184 we show that interfacing FFA into late model layers drives object recognition
185 comparably to the lateral occipital complex (fig 2B) on non-face natural images.
186 We suspect that the function of a region will only be fully understood by consid-
187 ering task-relevant variance across several tasks in light of activity in connected
188 brain regions. The tight interface we champion between computational models
189 and brain activity should prove useful in evaluating theoretical accounts of how
190 the brain solves tasks over time.

191 Computational models that perform the tasks end-to-end, from stimulus to
192 behaviour, should be particularly useful. In essence, translating between brain
193 regions to layers of such models can make clear what role a brain region plays
194 within the overall computation. In the case of object recognition, our results
195 suggested that recurrent models may be best positioned to explain how the
196 nature of information within brain regions changes as the computation unfolds.

197 This conclusion is in line with a growing body of modelling work in neuro-
198 science that affirms the value of recurrent computation(14, 15). Unlike the
199 aforementioned work, we suggest that long-distance recurrent connections that
200 link disparate layers should be considered (cf. (16)). We suspect such models
201 will be necessary to capture time course data and the duality found in some
202 brain regions, namely how most variance in a brain region can be attributable
203 to lower-level stimulus properties while co-mingled with important higher-level,
204 task-relevant signals.

205 As deep learning accounts in neuroscience are extended to other domains,
206 such as audition (17), and language processing (18), the lessons learned here
207 may apply. Our brain-model interface approach can help evaluate whether the
208 brain processes signals across domains in an analogous fashion. By minding the
209 distinction between shared and task-relevant variance, the role brain regions
210 play within the overall computation may more readily come into focus.

211 Our approach may also have practical application in brain machine inter-
212 faces (BMI). Recent BMI developments have emphasised the readout of motor
213 commands, neural processes taking place close to the periphery. In contrast,
214 by leveraging the constraints provided by a pre-trained DCNN, we were able to
215 gain traction on the ‘stuff of thought’, categorical and conceptual information
216 in IT. Because we learned a general translation from brain to model, our ap-
217 proach applied to BMI would allow distant generalisation. For example, we were
218 able to extrapolate to novel categories (see SI). For example, a translation from
219 brain to model that never trained on horses, but trained on other categories,
220 can perform zero-shot generalisation when given brain activity elicited by an
221 image of a horse. The interface has the potential to produce a domain-general
222 mapping rather than one dependent on specific training data. In the future,
223 BMI approaches that address general thought without exhaustive training on
224 all key elements and their combinations may be feasible.

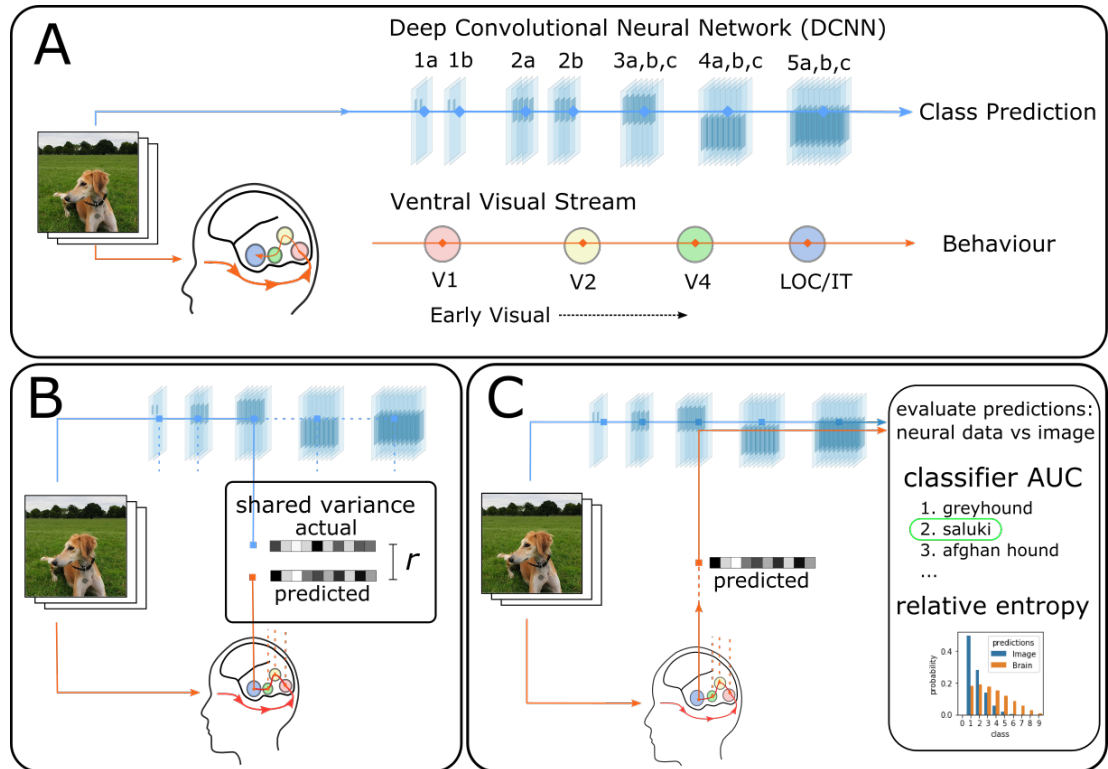


Figure 1: Deep Convolutional Neural Networks (DCNNs) trained on large naturalistic image datasets(19) have emerged as leading models of the mammalian ventral visual stream. **(A)** Typically, processing in DCNNs is hierarchical starting with the stimulus and proceeding across successive layers as higher-level information is extracted, culminating in predicting the class label(9). Numerous analyses(2-4) based on shared variance suggest the brain follows related principles with an early-to-early, late-to-late pattern of correspondence between the ventral visual stream and DCNN layers. **(B)** These shared-variance correspondences are evaluated locally, typically involving one brain region and one model layer, with no recourse to behaviour (i.e., the object recognition decision). **(C)** We propose a stronger test of correspondence based on task-relevant variance. If a model layer and brain region correspond, then model activity replaced with brain activity should drive the DCNN to an appropriate output (i.e., decision). The quality of correspondence is evaluated by comparing DCNN performance when driven by a stimulus image vs. interfaced with brain activity.

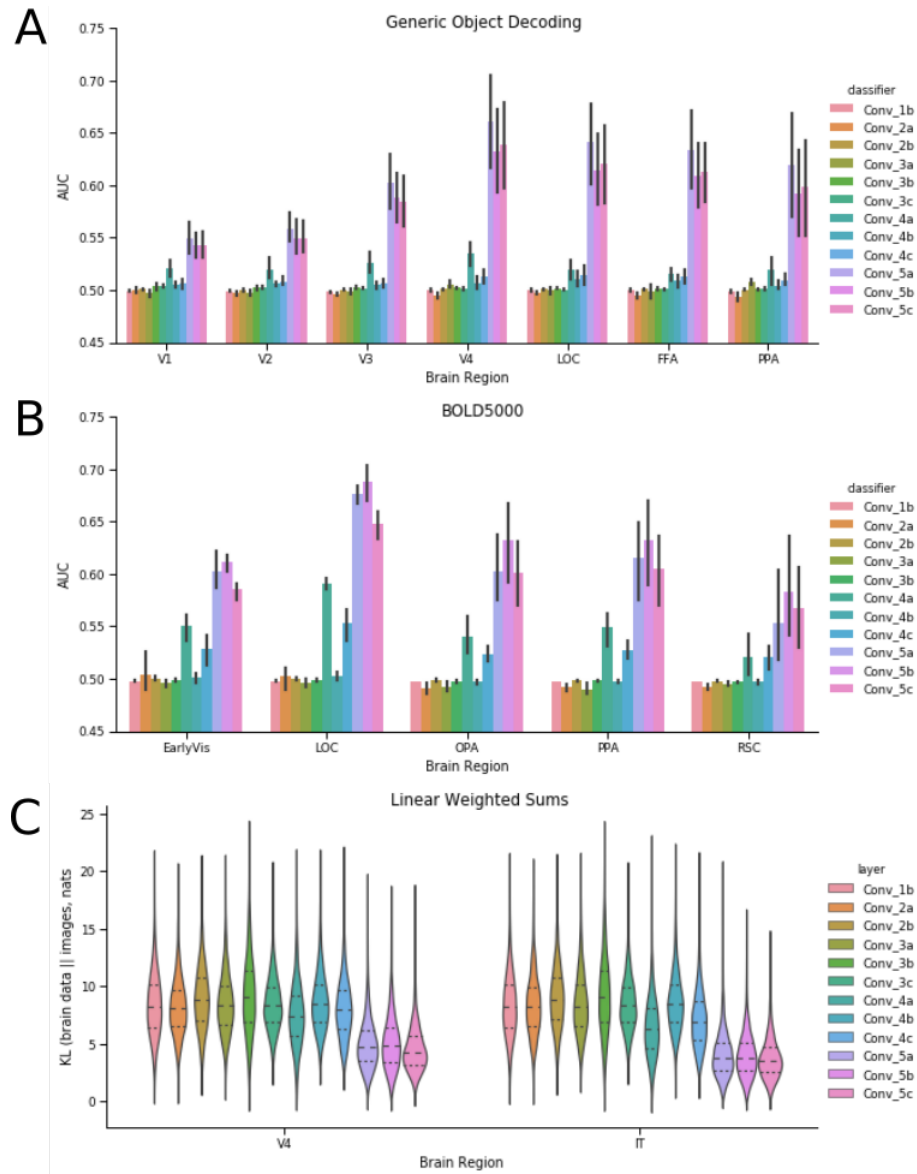


Figure 2: Results from interfacing neural data with a Deep Convolutional Neural Network (DCNN). Using the method shown in fig 1C, brain activity is directly inputted to a model layer to assess correspondence between a brain region and model layer. (A) For this human fMRI study(11), all brain areas drive DCNN object recognition performance to above chance levels. Performance is best for all brain areas when interfaced with later model layers. (B) The same pattern of results is found for a second human fMRI study(10). (C) In a third study, KL divergence is used (see main text and SI) to measure the degree of correspondence for when the DCNN is driven by image input vs. multi-unit recordings from macaque monkeys(12). For KL divergence, lower values indicate better correspondence. Once again, all regions best correspond to later network areas. These three analyses indicate that higher-level visual information is present at all stages along the ventral visual stream.

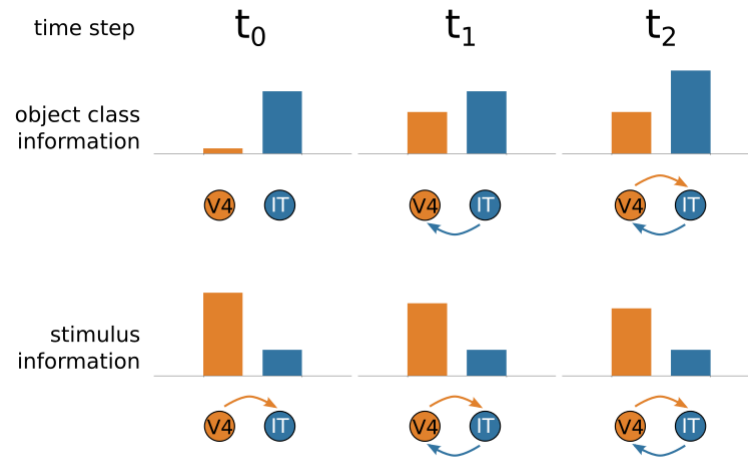


Figure 3: Hypothesised interactions between early (V4) and late (IT) regions along the ventral visual stream as processing unfolds. We hypothesise how stimulus and object-class information propagates between V4 and IT over time. At t_0 , the forward pass reaches IT from V4, with V4 activity reflecting low-level stimulus properties but little information about object class. At t_1 , object-class information from IT flows back to V4, increasing its task-relevant activity, which in turn influences IT at t_2 . Notice that later in processing, V4 reflects object class information, but most of its activity remains tied to bottom-up stimulus properties. These hypothesised interactions would reconcile our results (fig 2) based on task-relevant information with previous results based on shared variance.

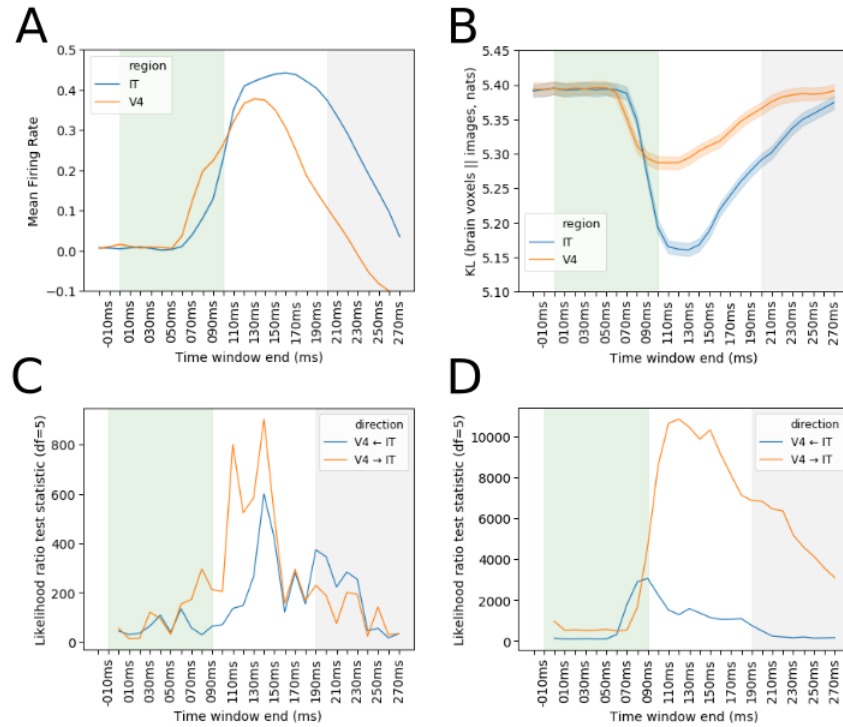


Figure 4: Analyses of monkey multi-unit recordings(12) time locked to stimulus presentation in 10ms time bins. Each visual stimulus was presented for 100ms (shaded green) with 100ms before the next (shaded grey). (A) Mean normalized spike counts for all electrodes for V4 and IT. (B) Task-relevant analysis (lower values imply closer correspondence with a late DCNN layer) show both V4 and IT can appropriately drive DCNN response (fig 1B), starting around 70ms after stimulus onset. (C) Consistent with our long-range recurrence hypothesis (fig 3), Granger Causal Modelling indicates that, while V4 first drives IT in terms of raw firing rates ($V4 \rightarrow IT$), (D) IT first drives V4 in terms of task-relevant information ($V4 \leftarrow IT$). These results are consistent with information about object category information (as assessed by interfacing with a late layer in a DCNN) first arising in IT and then feeding back to V4. At later time steps, Granger causality between V4 and IT becomes reciprocal ($V4 \leftrightarrow IT$) as the loop cycles.

225 References

- 226 1. Geirhos, R. *ImageNet-trained CNNs are biased towards texture; increasing*
227 *shape bias improves accuracy and robustness.* in (New Orleans, LA, USA,
228 2019). <https://openreview.net/forum?id=Bygh9j09KX>.
- 229 2. Güçlü, U. & van Gerven, M. A. Deep neural networks reveal a gradient in
230 the complexity of neural representations across the ventral stream. *Journal*
231 *of Neuroscience* **35**, 10005–10014 (2015).
- 232 3. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep Supervised, but Not Un-
233 supervised, Models May Explain IT Cortical Representation. *PLOS Com-*
234 *putational Biology* **10**, e1003915. ISSN: 1553-7358. (2019) (Nov. 2014).
- 235 4. Yamins, D. L. *et al.* Performance-optimized hierarchical models predict
236 neural responses in higher visual cortex. *Proceedings of the National Academy*
237 *of Sciences* **111**, 8619–8624 (2014).
- 238 5. Siegle, J. H. *et al.* Survey of spiking in the mouse visual system reveals
239 functional hierarchy. en. *Nature* **592**, 86–92. (2021) (Apr. 2021).
- 240 6. Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K.
241 Single-trial neural dynamics are dominated by richly varied movements.
242 *Nature Neuroscience* **22** (Oct. 2019).
- 243 7. Kragel, J. E., Morton, N. W. & Polyn, S. M. Neural activity in the medial
244 temporal lobe reveals the fidelity of mental time travel. *The Journal of*
245 *Neuroscience: The Official Journal of the Society for Neuroscience* **35**,
246 2914–2926 (Feb. 2015).
- 247 8. Purcell, B. A. *et al.* Neurally constrained modeling of perceptual decision
248 making. *Psychological Review* **117** (2010).
- 249 9. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for
250 Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*. arXiv: 1409.1556
251 (Sept. 2014).
- 252 10. Chang, N., Pyles, J. A., Gupta, A., Tarr, M. J. & Aminoff, E. M. BOLD5000:
253 A public fMRI dataset of 5000 images. *Scientific Data* **6** (2019).
- 254 11. Horikawa, T. & Kamitani, Y. Generic decoding of seen and imagined ob-
255 jects using hierarchical visual features. *Nature Communications* **8**, 1–15
256 (May 2017).
- 257 12. Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple learned
258 weighted sums of inferior temporal neuronal firing rates accurately predict
259 human core object recognition performance. *Journal of Neuroscience* **35**,
260 13402–13418 (2015).
- 261 13. McGugin, R. W., Gatenby, J. C., Gore, J. C. & Gauthier, I. High-resolution
262 imaging of expertise reveals reliable object selectivity in the fusiform face
263 area related to perceptual performance. *Proceedings of the National Academy*
264 *of Sciences* **109** (Oct. 2012).

- 265 14. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence
266 that recurrent circuits are critical to the ventral stream's execution of core
267 object recognition behavior. *Nature Neuroscience* **22**, 974 (2019).
- 268 15. Kubilius, J. *et al.* *Brain-Like Object Recognition with High-Performing*
269 *Shallow Recurrent ANNs in Advances in Neural Information Processing*
270 *Systems* **32** (2019).
- 271 16. Nayebi, A. *et al.* *Task-driven convolutional recurrent models of the vi-*
272 *sual system in Advances in Neural Information Processing Systems* (2018),
273 5290–5301.
- 274 17. Computational similarities between visual and auditory cortex studied
275 with convolutional neural networks, fMRI, and electrophysiology. *Journal*
276 *of Vision* **15**. (2021).
- 277 18. Schrimpf, M. *et al.* The neural architecture of language: Integrative reverse-
278 engineering converges on a model for predictive processing. *bioRxiv* (Oct.
279 2020).
- 280 19. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge.
281 en. *International Journal of Computer Vision* **115**, 211–252. [https://](https://doi.org/10.1007/s11263-015-0816-y)
282 doi.org/10.1007/s11263-015-0816-y (Dec. 2015).
- 283 20. Chollet, F. *keras* 2015. <https://github.com/fchollet/keras>.
- 284 21. Roads, B. D. & Love, B. C. Enriching ImageNet with Human Similarity
285 Judgments and Psychological Embeddings. *arXiv:2011.11015 [cs]* (Nov.
286 2020).
- 287 22. Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J. & Kriegeskorte,
288 N. Diverse deep neural networks all predict human IT well, after training
289 and fitting. *bioRxiv* (May 2020).
- 290 23. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network
291 Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]* (Feb.
292 2015).
- 293 24. Hastie, T. *The elements of statistical learning : data mining, inference,*
294 *and prediction /* (Springer, 2009).
- 295 25. Schrimpf, M. *et al.* Integrative Benchmarking to Advance Neurally Mecha-
296 nistic Models of Human Intelligence. en. *Neuron* **108**. (2021) (Nov. 2020).
- 297 26. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding Neural Rep-
298 resentational Spaces Using Multivariate Pattern Analysis. *Annual Review*
299 *of Neuroscience* **37**. (2019) (2014).
- 300 27. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal*
301 *of machine Learning research* **12**. Publisher: JMLR. org, 2825–2830 (2011).

302 5 Acknowledgements

303 The authors thank colleagues in the LoveLab for discussion and comments on
304 early versions of this manuscript. **Funding:** This research was supported by

305 NIH Grant 1P01HD080679 (<https://www.nih.gov/>), Royal Society Wolfson Fel-
306 lowship 183029 (<https://royalsociety.org/>), and a Wellcome Trust Senior Investi-
307 gator Award WT106931MA (<https://wellcome.org/>) held by B.C.L. The funders
308 had no role in study design, data collection and analysis, decision to publish,
309 or preparation of the manuscript. **Competing Interests:** The authors have
310 declared no competing interests exist. **Author Contributions:** N.J.S: Con-
311 ceptualization, methodology, software, validation, formal analysis, investigation,
312 data curation, writing - original draft, writing - review & editing, visualization.
313 B.C.L.: Conceptualization, methodology, resources, writing - review & editing,
314 supervision, funding acquisition.

315 6 Methods and Materials

316 6.1 Datasets

317 We re-analysed three existing neural datasets. Two, BOLD5000(10) and Generic
318 Object Decoding(11) consist of fMRI from human subjects who viewed images
319 taken from Imagenet(19), a benchmark large dataset of natural images. We
320 restricted the BOLD5000 dataset to only those images drawn from Imagenet
321 (2012 ILSVRC) edition and to subject 1-3 who completed the full experiment.
322 The analysis of Generic Object Decoding used the data from the ‘training’ por-
323 tion of their image presentation experiment, consisting of 1200 images from 150
324 categories drawn from the Imagenet Fall 2011 edition. For both datasets, each
325 image was presented once, thus each row represents individual trials.

326 The third dataset consists of neuron spike counts directly recorded from
327 V4 and IT of two macaque monkeys (12), in a rapid serial visual presentation
328 paradigm where each image is passively viewed for 100ms, with 100ms between
329 images. We used the publicly available data processed as detailed in those
330 publications. For the neural interfacing analysis of the spiking neural dataset,
331 we used spike rates aggregated over multiple presentations of each of 3200 unique
332 images, in the interval 70-170ms after stimulus onset, with the electrodes from
333 the two subjects concatenated, as in the original analysis (12). For the Granger
334 causal modelling analysis of the same dataset, we used spike rates at the level
335 of the individual trial (i.e., no aggregation) for each 10ms time bin.

336 The neural data corresponding with each image was related to layer activa-
337 tions of a deep convolutional neural network (DCNN) trained on image classi-
338 fication, when processing the same pixel-level data. The three neural datasets
339 contain data for various brain regions from ventral stream, including visual areas
340 (V1, V2, V3 V4, included as ‘EarlyVis’ in (10)), areas responsible for process-
341 ing shape and conceptual information (LOC, IT) and various downstream areas
342 (OPA, PPA, FFA, RSC).

343 For details on neuroanatomical placement or functional localisation of each
344 region, we refer readers to the original publications. Further details of brain
345 regions and dimensionality of the data from each region are presented in sup-
346plementary information table 1.

347 6.2 DCNN

348 As the base DCNN for all simulations, we used a re-implemented and trained
349 version of VGG-16 (9) (configuration D) using Keras(20) version 2.2.4 and
350 TensorFlow version 1.12. This model was selected for its uncomplicated archi-
351 tecture, near-human level classification accuracy on ImageNet, and widely
352 reported robust correspondence with primate or human data on various mea-
353 sures, including human behavioural (similarity judgements (21), human image
354 matching (15)) and neural (15, 22). We implemented and trained a version of
355 the architecture with $64 \times 64 \times 3$ input size, with corresponding changes in spatial
356 dimensions for all layers (table 2). For all analyses, images from all datasets were

357 cropped to a square and resized to this resolution. For the monkey multi-unit
358 dataset, where images are contained in a circular frame, the central 192×192
359 portion of the 256×256 original was cropped and resized, to decrease the pro-
360 portion of image taken up by blank space in the corners. While the original
361 authors trained their network in a two-stage process, beginning with a subset of
362 the layers, the inclusion of `batchnorm(23)` between the convolution operation
363 and activation function of each layer enabled training the complete network in a
364 single pass. We used the authors setting for weight decay (ℓ_2 penalty coefficient
365 of 5×10^{-4}) and a slightly different value for dropout probability (0.4). Model
366 architecture details are presented in the supplementary information (table 2).

367 **6.3 DCNN training**

368 Our training procedure followed (9). The model was trained on ImageNet 2012
369 (1000 classes) for analyses of the BOLD5000 and monkey multi-unit datasets.
370 For the Generic Object Decoding dataset, the model was trained to convergence
371 on ImageNet Fall 2011 (21841 classes), before layer FC3 was replaced and re-
372 trained with 150 classes, corresponding with the classes used in our re-analysis
373 of (11). For ImageNet Fall 2011 we randomly allocated 2% of each class in-
374 cluding all images used in (11) to an in-house validation set that was not used
375 for training. One image used by (11) was missing from our image dataset and
376 was excluded from all analyses. All images were resampled from their native
377 resolution to $64 \times 64 \times 3$ by rescaling the shortest side of the image to 64 pixels
378 and centre-cropping.

379 Both versions of the model was trained using mini-batch stochastic gradient
380 descent, with a batch size of 64, an initial learning rate of 0.001 and Nesterov
381 momentum of 0.90561. The learning rate decayed by a factor of 0.5 when
382 validation loss did not improve for 4 epochs, with training terminating after 10
383 epochs of no improvement. All layers used Glorot normal initialization. During
384 training, images were augmented with random rescaling, horizontal flips and
385 translations.

386 **6.4 Cross-validation**

387 Classifier-based methods require training classifier parameters, before evaluating
388 it on data withheld from the training set. In all analyses, we use the standard
389 approach of k -fold cross validation(24), in which the dataset is randomly allo-
390 cated into k equally-sized partitions, and the analysis is iterated k times, each
391 time training on $k - 1$ partitions and evaluating on one. In this way, the classifier
392 is evaluated over the entire dataset. For all analyses, except where otherwise
393 specified, we use stratified 8-fold cross validation, that is to say dataset items
394 are randomly allocated to partitions with the constraint that $1/k$ of each class
395 be allocated to each validation partition. For the spiking neural dataset(12),
396 each unique image was rendered from one of 64 objects, with varying position
397 and orientation. Here, stratification was done at the object level.

398 For the out-of-training-class generalisation analysis, we used leave-one-class-
399 out cross validation, where for m classes, the analysis is iterated m times, the
400 evaluation set consisting only of the entirety of a single class, on each iteration.

401 6.5 Neural Interfacing analysis

402 Given a dataset D , consisting of an image matrix D_i of shape $(n, 64, 64, 3)$ where
403 n is the number of images, and a corresponding neural data matrix D_r , of shape
404 (n, d) where d is the number of neural features (electrodes, for multi-unit data,
405 or voxels, for fMRI data), consider a DCNN computing a function f on D_i ,
406 mapping D to P_i , an (n, m) matrix of predictions, each row being a probability
407 distribution over the m classes the DCNN was originally trained to classify.

$$f(D_i) = P_i \quad (1)$$

408 For an arbitrary intermediate model layer q , we may decompose f into g_q and
409 g'_q , by computing intermediate activations, $g_q(D_i)$:

$$f(D_i) \equiv g'_q(g_q(D_i)) = P_i \quad (2)$$

410 The neural interface analyses proceeded by applying a linear transform W
411 to the centered and column-normalized neural data, D_r and inputting the result
412 into DCNN layer q , to compute a matrix of model predictions for the neural
413 data, P_r .

$$g'_q(WD_r) = P_r \quad (3)$$

414 6.5.1 Linear transformation matrix training

415 The transformation matrix W was computed by partitioning image and neu-
416 ral datasets D_i, D_r into training and evaluation partitions using 8-fold cross-
417 validation, and W was learned as a linear mapping from D_r to the layer q
418 activations generated by the corresponding images, D_i , on the training parti-
419 tion:

$$g_q(D_i) = WD_r + \epsilon \quad (4)$$

420 For each cross-validation fold, the model predictions were computed for the
421 evaluation partition. In practice, W was computed as a single-layer linear neural
422 network with no bias or activation function, to minimise mean-squared error of
423 supervision targets $g_q(D_i)$ using mini-batch stochastic gradient descent with
424 momentum, (batch size 64, momentum of 0.9, l_2 regularization of 0.0003, initial
425 learning rate of 0.1, decreasing by a factor of 0.5 when validation loss did not
426 improve for 4 epochs and terminating after 400 epochs or after validation loss
427 did not improve for 20 epochs.) For the analysis of the macaque dataset(12)
428 on the level of the individual trial, prior to performing the GCM model, W was
429 computed using the Adadelta optimizer (batch size of 128, initial learning rate
430 of 0.04.)

431 We also considered an alternative mode for training W , by first assembling
432 the model in the form of equation 3, composed of transformation matrix W
433 initialised with small random weights, followed by DCNN layer q onwards, g'_q ,
434 thus mapping end-to-end from neural measures D_r to output. W was then
435 trained by back-propagating the categorical cross-entropy error term from the
436 softmax output layer, using the supervision target of the ground-truth labels
437 for the neural dataset (D_r), with all other weights in the network frozen. This
438 method produced a pattern of results that was qualitatively similar, although
439 with lower absolute accuracy (SI fig 8).

440 6.5.2 Neural Interface Evaluation

441 The output of the model, P , is an (n, m) matrix of probability distributions
442 over the m output classes the original DCNN was trained on, for each of n
443 images in D . We computed this for the original DCNN on the image dataset,
444 $f(D_i) = P_i$, and also for the neural dataset for each brain region r and model
445 layer q , $g'_q(WD_r) = P_r$. The correspondence between r and q was evaluated
446 by comparing the model predictions P_r either against the ground-truth classes
447 (by computing the overall AUC of the classifier, via the equality between AUC
448 and Wilcoxon-Mann-Whitney U) or against model predictions from the image
449 dataset, by computing the KL divergence of P_r from P_i for each row n .

450 6.6 Shared Neural Variance Analysis

451 For comparison, we present an example of a shared neural variance analysis
452 using the macaque spiking neuron dataset (12) and our re-implemented model.
453 Conceptually, in common with the interfacing analysis (section 6.5), the analysis
454 evaluates the correspondence between a brain region r and a model layer q .
455 Layer q model activations, $g_q(D_i)$, were compared with a neural dataset obtained
456 from the presentation of corresponding images, D_r . To establish our results are
457 comparable to those previous, we used the neural predictivity method exactly
458 as implemented in the Brainscore benchmark for DCNNs (25).

459 The dataset was iteratively partitioned using 8-fold cross-validation into
460 training/validation partitions. Following the method of (25), we used the image
461 stimuli from the training partition to generate model activations on each layer.
462 We used PCA to calculate the first 1000 principal components of these activa-
463 tions, before training a PLS regression model (25 components) to predict, for
464 each electrode, the firing rate across the validation partition. The predictivity
465 for each electrode was computed as the Pearson correlation coefficient between
466 the predicted firing rates across the dataset and the actual recorded values,
467 with the overall predictivity given by the correlation coefficient of the median
468 electrode.

469 **6.7 Simple Classifiers on the Neural Datasets**

470 To establish performance baselines for the interfaced fMRI datasets, which were
471 evaluated in terms of classification performance, we applied various standard
472 classifiers to the neural data directly, to predict the image class from the neu-
473 ral data from various brain regions. Known as multi-voxel pattern analysis
474 (MVPA), evaluating the trained classifier’s ability to predict class labels from
475 fMRI or spiking neural data is now a standard approach to quantifying the
476 categorical-level information within a brain region (26). Nevertheless, in the
477 present analyses the number of different classes is unusually large, and the
478 number of examples from each class unusually small, (1916 images from 958
479 classes(10), 1200 images from 150 classes(11)) for a straightforward MVPA
480 analysis on these datasets. We report the AUC of the classifier computed in
481 the same way as for the neural interfacing analysis 6.5.2. All classifiers were
482 implemented as detailed below using version 0.20.3 of the Scikit-Learn library
483 (27).

484 **6.7.1 Multiclass Logistic Regression**

485 Implemented as LogisticRegression with the ‘multinomial’ option, the lbfgs
486 solver and a maximum of 10^3 iterations.

487 **6.7.2 Nearest Neighbours Classifier**

488 Implemented as KNeighboursClassifier. Given the structure of the BOLD5000
489 dataset, with only two examples per class (thus, either one or two examples in
490 the training partition, test classification of each class on the basis of one correct
491 training example) we classified on the basis of the single nearest neighbour under
492 a Euclidean distance function.

493 **6.7.3 Linear Support Vector Machine (SVM)**

494 Implemented as LinearSVC, using a one-versus-rest multi-class strategy, with a
495 maximum of 10^4 iterations and C parameter of 10^{-3} .

496 **6.8 Granger Causal Modelling**

497 In contrast to the previous neural interfacing analysis of the spiking neural
498 dataset, which aggregated spike rates over multiple presentations of each image,
499 in the interval 70-170ms after stimulus onset, here we trained and evaluated the
500 model on data at the individual trial level. We conducted a separate decoding
501 analysis for each 10ms time bin, from -20ms (i.e., prior to stimulus onset) to
502 270ms, with all time indices referring to the preceding 10ms bin. Training
503 linear transformation matrix W is described in section 6.5.1. Prior to the GCM,
504 we pre-processed the trial-level relative entropy data to ensure stationarity by,
505 first, subtracting the temporal mean and standard deviation from each trial,

506 and second, subtracting the mean signal and dividing by the signal's standard
507 deviation, thus ensuring that each time step has zero mean and unit variance.

508 Given two regions, X and Y, separate Granger-Causal models were computed
509 for each direction $X \rightarrow Y$ and $X \leftarrow Y$, where each model takes the form of a
510 linear regression, where the univariate outcome:

$$KL(D_X || D_i)_n \quad (5)$$

511 the KL divergence of region X with θ , the base model predictions, is predicted
512 by the Granger null model (6), or the Granger-causal model (7).

$$KL(D_X || D_i)_{n-1}, \dots, KL(D_X || D_i)_{n-p} \quad (6)$$

513

$$KL(D_X || D_i)_{n-1}, KL(D_Y || D_i)_{n-1}, \dots, KL(D_X || D_i)_{n-p}, KL(D_Y || D_i)_{n-p} \quad (7)$$

514 where p , the maximum number of previous time-steps is a hyperparameter
515 that is determined using model-selection criteria such as BIC. The appropriate
516 model was determined by comparing log-likelihood ratios, given the data, for
517 the causal and null models.

518 **A Appendix**

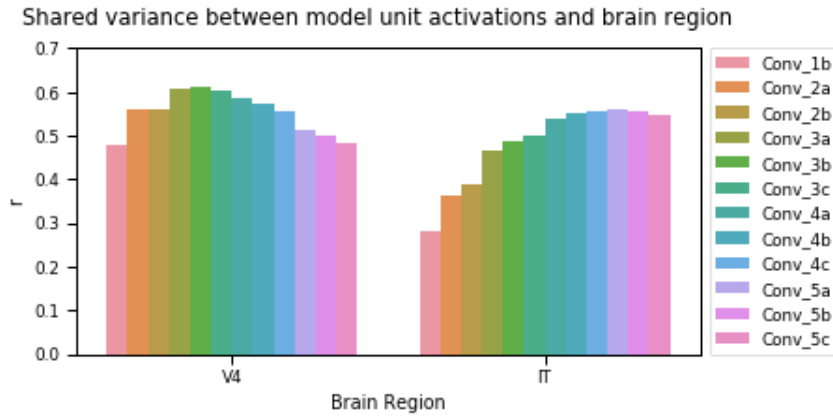


Figure 5: Standard approaches to relating primate ventral stream and DCNNs evaluate variance shared between data from each brain region and unit activations on each DCNN layer. They have been taken as evidence that earlier ventral stream regions (e.g., V4) correspond to earlier DCNN layers and later regions (e.g., IT) correspond to later DCNN layers. Here, we present a shared-variance based analyses of directly recorded spiking neural activity(12) and VGG-16 using an established method (15). Higher correlations reflect more shared variance between brain region and model layer.

519 **B Supplementary Figures**

520 **C Tables**

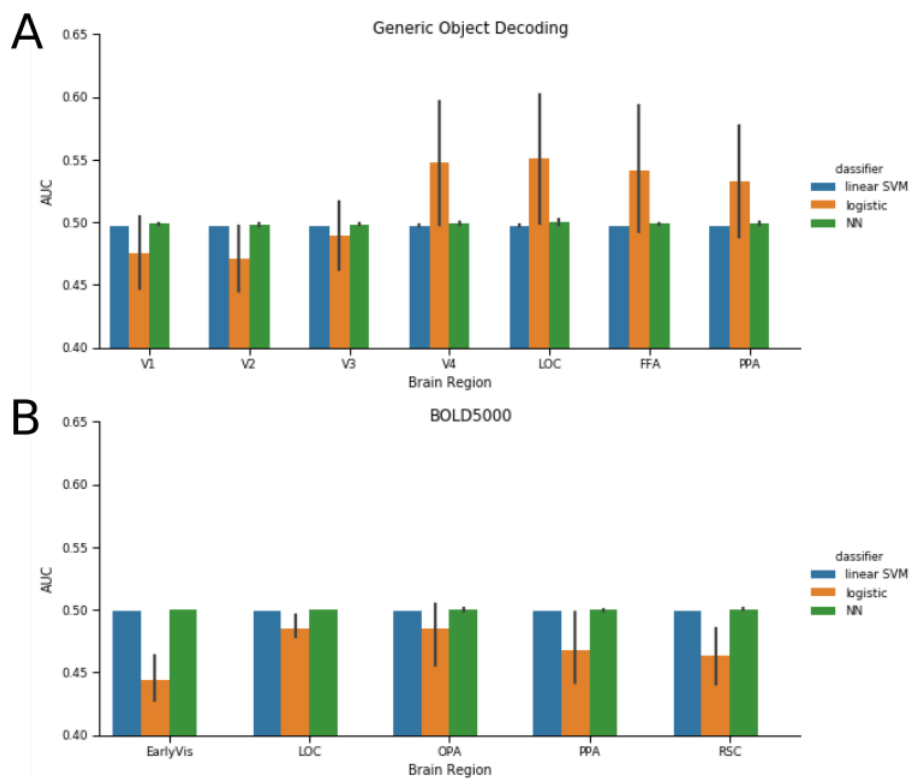


Figure 6: Comparison of the brain-DCNN interface as a neural pattern classifier, compared with standard linear classifiers typically used in multi-voxel pattern analysis (MVPA). We present classification performance (AUC) directly on neural patterns, on the Generic Object Decoding (A) and BOLD5000 (B) datasets, for simple classifiers (support vector machine with linear kernel, multiclass logistic regression, and a 1-nearest neighbour classifier) with results for interface with each layer of the DCNN presented for comparison. Performance of the simple classifiers is generally near chance (0.5), we attribute this to the large number of image classes (150, 958 respectively) and few available examples (2, 8 per class) which severely limit the available training data. Because the brain-DCNN interface learns a general mapping between brain region and model, it does not suffer this limitation, making it an appealing novel approach for MVPA.

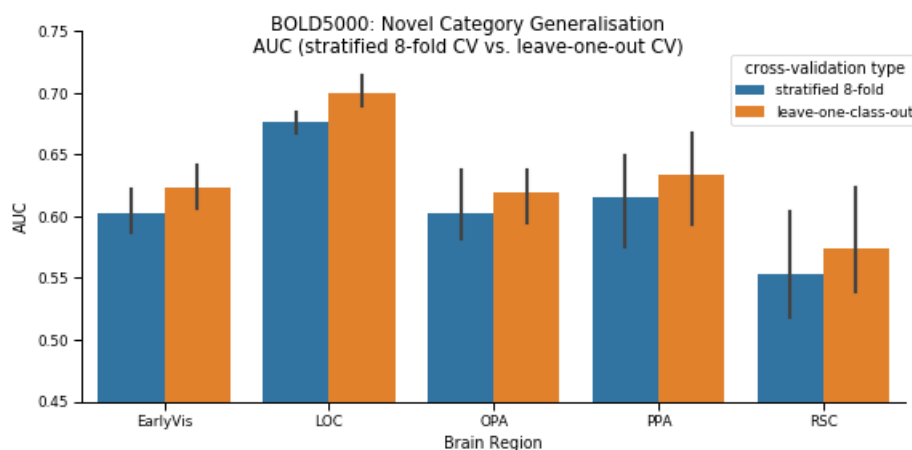


Figure 7: Learning a mapping directly from neural measures to DCNN activation space produces a general mapping, rather than being dependent on training examples. The neural interface has an in-built ability to generalise to novel classes. This is demonstrated by presenting classification performance (AUC) on the BOLD5000 dataset, by comparing cross-validation (CV) strategies. Error bars represent 95% confidence intervals across 3 subjects. Stratified 8-fold CV (the default, used in all other analysis) ensures each training partition contains at least one example of each class. Leave-one-class-out CV involves the same number of CV folds as there are classes, each time training on all data except one class, which is withheld for the validation set. Performance is equivalent or better (LOC) when generalising to novel classes, which we attribute to more training data per CV fold. Due to the training time, this analysis was restricted to layer 5a.

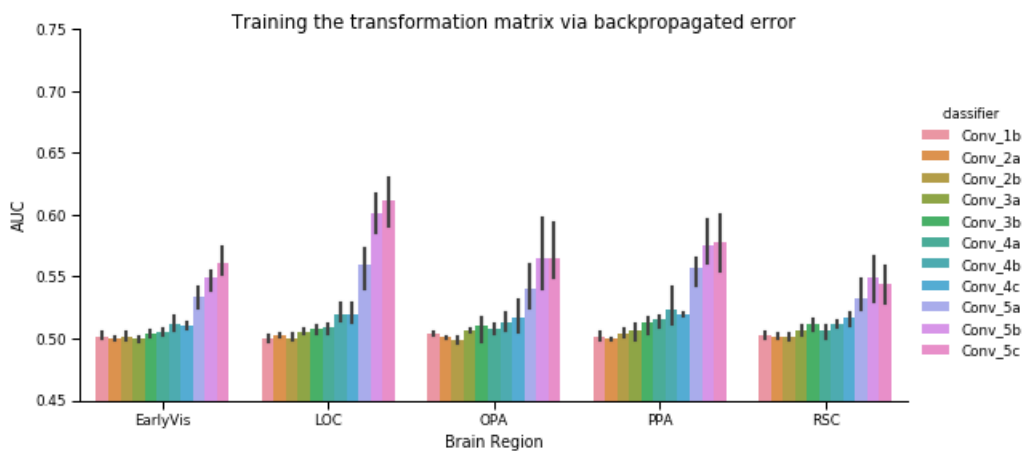


Figure 8: Alternative ‘backprop mode’ for training the transformation matrix W mapping from neural space to DCNN activation space. Classification performance (AUC) on the BOLD5000 dataset follows a qualitatively similar pattern to the main analysis (compare figure 2B), albeit with lower absolute accuracy. The default analysis trains W independently as a regression problem, using layer activations as supervision targets directly. Instead, this approach uses W as a weights matrix for a new neural network that takes neural data from a brain region as input, connected to the latter part of the DCNN, and training the network using the class labels as supervision targets, with all other DCNN weights frozen.

Dataset	Generic Object Decoding(<i>11</i>)	BOLD5000(<i>10</i>)	Linear Weighted Sums(<i>12</i>)
Stimuli	experiment ‘train’ phase: 1200 images from 150 categories (ImageNet Fall 2011)	1916 images from 958 categories (ImageNet ILSVRC 2012)	3200 greyscale composite images, 64 objects in 8 categories, non-congruent background
Task	one-back repetition detection	valence judgement (‘like’, ‘neutral’, ‘dislike’)	passive viewing, RSVP presentation 100ms/100ms
Subjects	5 human fMRI	3 human fMRI (partial data from subject 4 excluded)	2 Macaque monkeys (vectors concatenated) multi-unit recording
Time indices	full 9s of image presentation	TR3-4	70-170ms
Brain region (dimensionality per subject)	V1 (1004, 757, 872, 719, 659) V2 (1018, 944, 1031, 855, 891) V3 (759, 810, 861, 929, 907) V4 (740, 544, 754, 704, 860) LOC (540, 834, 996, 668, 566) PPA (356, 316, 496, 398, 550) FFA (568, 435, 928, 725, 929)	EarlyVis (495, 495, 1218) LOC (342, 888, 1027) OPA (288, 180, 392) PPA (331, 370, 273) RSC (229, 421, 394)	IT (168 = 58 + 110) V4 (88 = 70 + 18)

Table 1: **Neural datasets** For further dataset details, such as how regions were defined, we refer readers to the original publications

Block	Layer	Dimensions ($h \times w \times c$)	Filter Size
Input		$64 \times 64 \times 3$	
1	1a	$64 \times 64 \times 64$	3×3
	1b	$64 \times 64 \times 64$	3×3
	max pool 1		2×2
2	2a	$32 \times 32 \times 128$	3×3
	2b	$32 \times 32 \times 128$	3×3
	max pool 2		2×2
3	3a	$16 \times 16 \times 256$	3×3
	3b	$16 \times 16 \times 256$	3×3
	3c	$16 \times 16 \times 256$	3×3
	max pool 3		2×2
4	4a	$8 \times 8 \times 512$	3×3
	4b	$8 \times 8 \times 512$	3×3
	4c	$8 \times 8 \times 512$	3×3
	max pool 4		2×2
5	5a	$4 \times 4 \times 512$	3×3
	5b	$4 \times 4 \times 512$	3×3
	5c	$4 \times 4 \times 512$	3×3
	max pool 5		2×2
FC	FC1	4096	
	dropout 1		
	FC2	4096	
	dropout 2		
	FC3 (output)	1000 softmax	

Table 2: **DCNN Architecture:** Layer configuration and dimensions of the DCNN used for all analyses.