nature portfolio

Peer Review File

A Large-Scale Examination of Inductive Biases Shaping High-Level Visual Representation in Brains and Machines

Reviewer #1 (Remarks to the Author):

The authors have carried out a set of controlled experiments that compared different DNNs that vary in theoretically relevant ways. They find the training dataset is the most important variable in accounting for RSA predictions.

This is one of the few research projects that carry out controlled experiments to determine which factors contribute to good brain predictivity in DNNs. The authors have done an impressive job comparing many different models, and the message is straightforward and interesting. However, there are also some important limitations of the experiments that weaken or undermine key conclusions that are drawn. I think these issues should be addressed or at least explicitly discussed in a revision.

The key conclusion is the following: "These results indirectly reveal a currently unquantified factor of dataset diversity as an important predictor of brain-like visual representation". That is, they are claiming that good RSAs are the product "brain-like visual representation".

But this standard logic is not justified– correlations between representational geometries does not mean that the two systems encode knowledge in a similar way. Indeed, the systems may not even be encoding the same visual features, with high RSAs reflecting confounds in datasets (Dujmović et al., 2022). When researchers have carried out experiments that systematically manipulate the images rather than the models, it becomes clear that many DNNs are classifying images largely based on texture (Geirhos et al., 2019), and these texture representations support good brain-scores and RSA scores. By contrast, humans largely rely on shape when identifying objects. Almost certainly this is the case with most or all the models tested here – with texture confounded with shape. How does that impact on the conclusions the authors want to draw? The authors should at least discuss the possibility that the high RSAs scores reflect learned texture representations that are correlated (confounded) with shape, and that these studies are not providing evidence that DNNs are learning more human-like object representations when trained on a more diverse set of object categories.

The main claim is that it is the diet of training images that lead to better predictions. The authors should discuss other studies that observe that untrained models sometimes produce similar RSA scores. For instance, Storrs et al. (2021) write:

"we compared the hIT correlation of every layer in trained and untrained versions of each model (Figure 3). We found that training improved representational similarity to hIT, but by a perhaps surprisingly small degree…whereas untrained models showed similar hIT correlation across all their layers, the performance of trained models peaked for processing steps about one half to three fourths of the way from network input to output".

These findings should be noted and some attempt to reconcile the current findings with previous findings should be provided.

The authors should emphasize more (e.g., note in the figure captions) that the RSA scores reflect the layer that predicts the brain response best. In Storrs et al., the best predictions occurred in the middle layers of networks. Is that the case here? What are we to make if the difference between trained and untrained models is not so great in later layers?

Given the main claim is that training that leads to better performance, and given the contrasting findings from some previous work, perhaps it would be good to more systematically study RSA in untrained networks. For example, in Figure 1, could untrained models be included for all models and the results included in the Figure?

Similarly, the authors should cite and discuss the findings of Xu and Vaziri-Pashkam (2021) who reported RSA scores are much reduced for novel stimuli. This is another study that manipulated the images in ways that seems to undermine the claim that higher RSA Scores reflect similar brain like representations – if they were similar, the RSA scores would not plummet for novel objects. How can the authors reconcile this finding with the conclusions they want to draw?

I’m not sure I agree (or understand) the authors characterization of neuroconnectionism when they write:

“..we conceptualize these models less as in silico models of the brain with one-to-one correspondence to different regions, and more as abstracted visual representation learners, with representational signatures that are either more or less akin to the biological visual system. …”

This seems at odds with claims such as:

“The empirical reason why ANNs can be called the "current best" models of human vision is that they offer unprecedented mechanistic explanations of the human capacity to make sense of complex, naturalistic inputs”. (Golan et all, in press, BBS response).

Of course, neuroconnectionist models (like all models) are more abstract than the phenomena they model, but it seems to me that DNN models of human vision are models of brain that support vision, no? Perhaps some clarification here would be useful, as I’m not sure what “visual representation learners, with representational signatures” of the biological visual system means.

I also think the following somewhat mischaracterizes how neuroconnectionism works in general:

“Within this framework, training sets of models that vary only in one of these factors, while controlling all others, can be thought of as controlled rearing experiments (Wood et al., 2020), operationalizing different artificial visual systems to explore targeted questions about which variations give rise to a more or less emergent brain-like representation.”

This is what the authors are doing here, but it is the exception. In the majority of cases, RSAs are compared across models that vary along multiple dimensions (as is also the case with the models in Brain-Score assessed with linear regression).

I appreciate the authors cited the Bowers et al BBS paper, but I think this paper should be cited in this context, as the key message of the paper is that DNNs need to be assessed on experiments in which independent variables are manipulated to test specific hypotheses. For example, from the abstract:

“More generally, theorists need to build models that explain the results of experiments that manipulate independent variables designed to test hypotheses rather than compete on making the best predictions.”

The BBS paper, not the neuroconnectionist approach, emphasizes this key methodological point that the authors are using here.

Signed: Jeff Bowers

Dujmović, M., Bowers, J. S., Adolfi, F., & Malhotra, G. (2022). Obstacles to inferring mechanistic similarity using Representational Similarity Analysis. bioRxiv, 2022-04.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. Journal of cognitive neuroscience, 33(10), 2044-2064.

Reviewer #2 (Remarks to the Author):

An excellent paper which makes an important contribution to the field, by going beyond leaderboards to ask "what, if anything, makes one DNN better or worse than another at predicting neural responses?" The question is an obvious one, which has been addressed in more limited ways by earlier papers (e.g. Storrs et al. 2021, Zhuang et al. 2021, others), but never at this scale

nor with this resolution and quantity of human fMRI data. The manuscript is very easy to read for a potentially technically dense paper, and beautifully illustrated.

MAIN COMMENTS

It would be good to clarify in text why this sort of analysis (systematically comparing groups of models varying factors like training objective or dataset) can't be done *within* some of the existing leaderboard competitions like BrainScore. Rather than being an *alternative* to finding the single best model, it seems like ideally these two quests could be combined within a single publicly-accessible and flexibly-searchable database: one could submit models to a project like BrainScore, their neural predictivity would be calculated, *and* one could then group the model results in various different ways in order to gauge the effect of training objective, training dataset, etc. Rather than pitching the work as fundamentally an alternative to leaderboard type comparisons, it would be helpful to spell out exactly what information is missing from current leaderboards that would allow one to do this sort of analysis.

MINOR COMMENTS

I appreciate the merit in leaving most technical details to the Methods. One thing I would like to see mentioned earlier on in the Results section though (e.g. Figure 2 caption and nearby) is that the "model performance" is the performance of the best single layer (evaluated non-circularly by using separate sets of data to identify the best layer and then measure its performance).

In the Feature Extraction section, some missing detail on how were images input to networks, e.g. presumably images were scaled to the input resolution of the network and perhaps in some cases normalised by subtracting the average of that network's training images. How much did these pre-input image processing steps differ across networks?

Reviewer #3 (Remarks to the Author):

In this manuscript the authors present a thorough analysis comparing deep neural network predictions of fMRI data. Concretely, they compare a large number of neural networks in how well they predict responses from the natural scenes dataset based on a representational similarity analysis (RSA). By selecting subsets of the DNN models that differ as specifically as possible in identifiable factors, they test a broad selection of hypotheses what might cause differences in model prediction quality and thus what might be necessary for a good model of the human visual cortex. Many of the factors thought to influence brain prediction quality turn out to be less informative than we might have hoped, in particular when a voxel encoding model is first trained to adjust the models predictions to the data, after which most models perform within 0.1 correlation from the best model. Nonetheless, there are some interpretable differences between models.

In general, I found the presented analyses convincing and very informative. The authors' comparisons along the different factors of variation are much more informative than the search for the ultimately best model. Also, the collection of models and hypotheses seems quite complete and serves as a review of the literature on deep neural networks as models of the human brain as well.

I have two points that I think the authors should address in this paper though:

1) All conclusions in this manuscript are drawn based on mean statistics among the compared groups, which weighs the evidence for all models equally. In particular, models that perform very badly are weighted the same way as the highest performing models. Arguably, this is not exactly what we want. Models that are bad models of the human visual system are also less representative for it such that we should weigh what helps for these bad models less than what happens for ones that otherwise perform similar to humans. I believe the authors follow this logic when they exclude the untrained models from the statistics in Figure 5. I don't think the conclusions in the manuscript are wrong, but think an argument should be made why the mean is

good enough here.

2) The authors are using well motivated single subject based statistics. I think the authors should defend that choice in comparison to statistics based on the group of subjects. There is always ongoing debate on how to best do the analyses for models of fMRI data fairly and how far the results can generalise based on which statistical analyses. Thus, I think a proper defence of the choices made is necessary. The authors do comment on some of these questions already in the supplement, but I believe some discussion in the main text is necessary.

Minor points:

As only 4 subjects entered the analyses, presenting those 4 datapoints in some appendix would be a good idea. Even if all the model comparisons were shown this would be a completely readable length.

I think a stronger separation of the different analyses in the plots would be great, i.e. not connecting the boxes with lines above the plots and a little more space between the columns. Perhaps a letter label for the parts. I certainly struggled to keep them apart and in most of them at least, there is space.

Response to Reviews (Point-by-Point)

We are deeply grateful to all reviewers for their engagement with our work, and their rigorous, comprehensive feedback. Our point-by-point responses to this feedback may be found below.

Original reviewer responses are formatted in black.

- Our responses are provided as bullet points in blue.

Response to Reviewer #1

The authors have carried out a set of controlled experiments that compared different DNNs that vary in theoretically relevant ways. They find the training dataset is the most important variable in accounting for RSA predictions.

This is one of the few research projects that carry out controlled experiments to determine which factors contribute to good brain predictivity in DNNs. The authors have done an impressive job comparing many different models, and the message is straightforward and interesting. However, there are also some important limitations of the experiments that weaken or undermine key conclusions that are drawn. I think these issues should be addressed or at least explicitly discussed in a revision.

- We thank the reviewer for this conceptualization of our work, and hope that various additions we have made to our manuscript, as well as the point-by-point responses below, will be sufficient to address the reviewer's concerns regarding potential limitations or weaknesses.

The key conclusion is the following: "These results indirectly reveal a currently unquantified factor of dataset diversity as an important predictor of brain-like visual representation". That is, they are claiming that good RSAs are the product "brain-like visual representation".

But this standard logic is not justified– correlations between representational geometries does not mean that the two systems encode knowledge in a similar way. Indeed, the systems may not even be encoding the same visual features, with high RSAs reflecting confounds in datasets (Dujmović et al., 2022). When researchers have carried out experiments that systematically manipulate the images rather than the models, it becomes clear that many DNNs are classifying images largely based on texture (Geirhos et al., 2019), and these texture representations support good brain-scores and RSA scores. By contrast, humans largely rely on shape when identifying objects. Almost certainly this is the case with most or all the models tested here – with texture confounded with shape. How does that impact on the conclusions the authors want to draw? The authors should at least discuss the possibility that the high RSAs scores reflect learned texture representations that are correlated (confounded) with shape, and that these studies are not providing evidence that DNNs are learning more human-like object representations when trained on a more diverse set of object categories.

- Thank you for raising this point. We have now added substantive new text to address the limitations of inference from high RSA values on whether a system is "brain-like" (see Discussion, lines 565-592). Specifically, we discuss the choice of both the stimuli used to probe the representations and the distance metrics used to

compare them as key analytical choices that matter for the kind of answers we get. And, that finding 'brain-predictive' scores here does not directly imply 'brain-like' representation, which we agree will be more fully revealed with more targeted tests (e.g. using Geirhos-style texture vs. shape biased stimuli among others, as suggested in the BBS paper).

The main claim is that it is the diet of training images that lead to better predictions. The authors should discuss other studies that observe that untrained models sometimes produce similar RSA scores. For instance, Storrs et al. (2021) write:

"we compared the hIT correlation of every layer in trained and untrained versions of each model (Figure 3). We found that training improved representational similarity to hIT, but by a perhaps surprisingly small degree…whereas untrained models showed similar hIT correlation across all their layers, the performance of trained models peaked for processing steps about one half to three fourths of the way from network input to output".

These findings should be noted and some attempt to reconcile the current findings with previous findings should be provided.

- In the revised Results section comparing trained vs. untrained models (lines 338-342), we now reference all the papers we could find where untrained and trained models yield similar brain predictivity (e.g. Cadena et al., 2019 (in mice); and Storrs et al., 2021), and further reference work by Hermann et al., 2020 and Baek et al. 2021, which examine the properties of untrained feature spaces. As counterpoint, we also cite a handful of studies that show an advantage for trained models over untrained models (e.g. Murty et al. 2021, Prince et al. 2023, Nonaka et al., 2021, etc., see lines 342-344)
- It is true that in Figure 3 of the Storrs et al., 2021 paper the trained vs untrained models are similar across some layers, when using the cRSA metric. However, we note that they were working with the rather limited data available at the time (brain responses to only 62 isolated objects to fit/predict, with relatively low noise ceilings from the early days of condition-rich fMRI design protocols; r=0.33-0.48). Our cRSA results encompass the same DNN models they used, but at substantially larger scale and with more reliable data collected with more powerful fMRI protocols (1000 images, noise ceiling r=0.71-0.86), and we find a qualitatively different pattern where trained models out-predict untrained models substantially. We thus suspect that differences in data quality may explain the divergence in results.

The authors should emphasize more (e.g., note in the figure captions) that the RSA scores reflect the layer that predicts the brain response best.

- Done! A note on this point has been added to the captions of Figures 2-6 and Supplementary Figure 2.

In Storrs et al., the best predictions occurred in the middle layers of networks. Is that the case here? What are we to make if the difference between trained and untrained models is not so great in later layers?

- In our data, on average the best fitting layer was at ~85% depth. However, (unlike in Storrs et al., 2021), we find that the difference between trained and untrained models is maintained across all the layers we tested through the late stages. Thus this potential concern does not apply here.
- We have now included a new supplementary figure that shows the brain prediction fits as a function layer depth, for trained and untrained architecture, and for both the cRSA and veRSA metrics (see Section SI.4, lines 1244-1255, Supplementary Figure 3).

Given the main claim is that training leads to better performance, and given the contrasting findings from some previous work, perhaps it would be good to more systematically study RSA in untrained networks. For example, in Figure 1, could untrained models be included for all models and the results included in the Figure?

- We agree that further scrutiny of the untrained versions of each architecture is informative, but have not added this to Figure 1, because it adds an additional interaction term (metric x architecture x *training*) that renders the plot too complex. Instead, we have added plots of trained versus untrained models, by layer, to the aforementioned supplement (see Section SI.4 and the new Supplementary Figure 3, mentioned in the main text at line 346).

Similarly, the authors should cite and discuss the findings of Xu and Vaziri-Pashkam (2021) who reported RSA scores are much reduced for novel stimuli. This is another study that manipulated the images in ways that seems to undermine the claim that higher RSA Scores reflect similar brain like representations – if they were similar, the RSA scores would not plummet for novel objects. How can the authors reconcile this finding with the conclusions they want to draw?

- We have added the citation of Xu and Vaziri-Pashkam (2021) in the revised Discussion section (line 574), where we consider the limitations that apply due to the stimuli selected for comparison between brains and models.

- Indeed, part of our main take home is that the choice of linking methods, including the chosen stimulus set, matters quite a bit for claims about whether or not these models are 'good' models of OTC representation.
- We have added a new paragraph in the general discussion that explicitly qualifies the scope of the inferences that our paper can make based on these choices (line 583-592).

I'm not sure I agree (or understand) the authors characterization of neuroconnectionism when they write:

"..we conceptualize these models less as in silico models of the brain with one-to-one correspondence to different regions, and more as abstracted visual representation learners, with representational signatures that are either more or less akin to the biological visual system. …"

This seems at odds with claims such as:

"The empirical reason why ANNs can be called the "current best" models of human vision is that they offer unprecedented mechanistic explanations of the human capacity to make sense of complex, naturalistic inputs". (Golan et all, in press, BBS response).

Of course, neuroconnectionist models (like all models) are more abstract than the phenomena they model, but it seems to me that DNN models of human vision are models of brain that support vision, no? Perhaps some clarification here would be useful, as I'm not sure what "visual representation learners, with representational signatures" of the biological visual system means.

- Thank you for highlighting this important conceptual point. We think of these deep neural networks as "model organisms" whose visual systems we can study in the same way we might study the mouse or monkey visual system. Studying these parallel systems allows us to explore how their variation, and emergent brain-predictive properties, reveal deeper principles that might also apply to the human visual system. At the same time, we also acknowledge there's not an intended 1-1 correspondence (mice and monkey brains are not human brains, but their visual systems might share common principles). We have now modified this part of the Introduction to clarify this argument (see lines 70-81).

I also think the following somewhat mischaracterizes how neuroconnectionism works in general:

"Within this framework, training sets of models that vary only in one of these factors, while controlling all others, can be thought of as controlled rearing experiments (Wood et al., 2020), operationalizing different artificial visual systems to explore targeted questions about which variations give rise to a more or less emergent brain-like representation."

This is what the authors are doing here, but it is the exception. In the majority of cases, RSAs are compared across models that vary along multiple dimensions (as is also the case with the models in Brain-Score assessed with linear regression).

- We have removed the reference to the Neuroconnectionist research paper in this part of the manuscript (see Introduction, lines 70-81), in an attempt to make clear that our research programme is not fully synonymous with the Neuroconnectionist research programme (though we think it is at least compatible with it!).
- And, we very much agree with the reviewer that this approach is perhaps less common than it should be! Our hope is that his paper will help to inspire more controlled deep net experimentation.

I appreciate the authors cited the Bowers et al BBS paper, but I think this paper should be cited in this context, as the key message of the paper is that DNNs need to be assessed on experiments in which independent variables are manipulated to test specific hypotheses. For example, from the abstract:

"More generally, theorists need to build models that explain the results of experiments that manipulate independent variables designed to test hypotheses rather than compete on making the best predictions."

The BBS paper, not the neuroconnectionist approach, emphasizes this key methodological point that the authors are using here.

- In our revised Introduction (lines 70-81) – where we situate how our modeling approach is different from current benchmarking paradigms – we no longer emphasize the Neuroconnectionist approach quite as strongly. Instead, we simply define our overarching framework and assumptions explicitly, without tying them to other frameworks whose explicit or latent assumptions and logic may not fully align with our own.
- Simultaneously, we have moved the bulk of comparison to other frameworks to our revised Discussion section, where we also expand more deeply now on the issues raised in your BBS paper, and cite it there in context (lines 566-578).

Signed: Jeff Bowers

Dujmović, M., Bowers, J. S., Adolfi, F., & Malhotra, G. (2022). Obstacles to inferring mechanistic similarity using Representational Similarity Analysis. bioRxiv, 2022-04.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. Journal of cognitive neuroscience, 33(10), 2044-2064.

Response to Reviewer #2

An excellent paper which makes an important contribution to the field, by going beyond leaderboards to ask "what, if anything, makes one DNN better or worse than another at predicting neural responses?" The question is an obvious one, which has been addressed in more limited ways by earlier papers (e.g. Storrs et al. 2021, Zhuang et al. 2021, others), but never at this scale nor with this resolution and quantity of human fMRI data. The manuscript is very easy to read for a potentially technically dense paper, and beautifully illustrated.

- Thank you for this kind feedback! We appreciate it.

MAIN COMMENTS

It would be good to clarify in text why this sort of analysis (systematically comparing groups of models varying factors like training objective or dataset) can't be done *within* some of the existing leaderboard competitions like BrainScore. Rather than being an *alternative* to finding the single best model, it seems like ideally these two quests could be combined within a single publicly-accessible and flexibly-searchable database: one could submit models to a project like BrainScore, their neural predictivity would be calculated, *and* one could then group the model results in various different ways in order to gauge the effect of training objective, training dataset, etc. Rather than pitching the work as fundamentally an alternative to leaderboard type comparisons, it would be helpful to spell out exactly what information is missing from current leaderboards that would allow one to do this sort of analysis.

- We have softened our use of the word “alternative”, emphasizing that our approach is also “complementary” (see Introduction, line 70)
- We also agree with the point about complementarity more generally, and see no *a priori* reason analyses like ours could not be done with current benchmarking platforms. The key addition required of these platforms or pipelines to allow for analyses like ours is simply a careful and comprehensive collection of model metadata, to serve as the basis of statistical grouping operations. We have included new text to clarify this on line 598.

MINOR COMMENTS

I appreciate the merit in leaving most technical details to the Methods. One thing I would like to see mentioned earlier on in the Results section though (e.g. Figure 2 caption and nearby) is that the "model performance" is the performance of the best single layer (evaluated non-circularly by using separate sets of data to identify the best layer and then measure its performance).

- We agree and have added this to the beginning of the results section (lines 135-140).

In the Feature Extraction section, some missing detail on how were images input to networks, e.g. presumably images were scaled to the input resolution of the network and perhaps in some cases normalised by subtracting the average of that network's training images. How much did these pre-input image processing steps differ across networks?

- We used the same image resolution and pre-transforms used to evaluate the models, based on the code repositories from which they were collected (directly porting the relevant code where possible). We now report these details in the Methods section (see lines 707-712). These transforms almost always involve a resize operation and normalization by the mean and (for pretrained models) standardized deviation of values in the training set (e.g. ImageNet).

Response to Reviewer #3

In this manuscript the authors present a thorough analysis comparing deep neural network predictions of fMRI data. Concretely, they compare a large number of neural networks in how well they predict responses from the natural scenes dataset based on a representational similarity analysis (RSA). By selecting subsets of the DNN models that differ as specifically as possible in identifiable factors, they test a broad selection of hypotheses what might cause differences in model prediction quality and thus what might be necessary for a good model of the human visual cortex. Many of the factors thought to influence brain prediction quality turn out to be less informative than we might have hoped, in particular when a voxel encoding model is first trained to adjust the models predictions to the data, after which most models perform within 0.1 correlation from the best model. Nonetheless, there are some interpretable differences between models.

In general, I found the presented analyses convincing and very informative. The authors' comparisons along the different factors of variation are much more informative than the search for the ultimately best model. Also, the collection of models and hypotheses seems quite complete and serves as a review of the literature on deep neural networks as models of the human brain as well.

- We thank the reviewer for this positive feedback.

I have two points that I think the authors should address in this paper though:

1) All conclusions in this manuscript are drawn based on mean statistics among the compared groups, which weighs the evidence for all models equally. In particular, models that perform very badly are weighted the same way as the highest performing models. Arguably, this is not exactly what we want. Models that are bad models of the human visual system are also less representative for it, such that we should weigh what helps for these bad models less than what happens for ones that otherwise perform similar to humans. I believe the authors follow this logic when they exclude the untrained models from the statistics in Figure 5. I don't think the conclusions in the manuscript are wrong, but think an argument should be made why the mean is good enough here.

- Thank you for your comments. We have made several clarifying changes throughout the manuscript, separately for our mean-based analyses (Fig. 2,3,4) and our correlation-based continuous value analyses (Fig, 5).
- First, our primary analyses draw on mean statistics (Fig 2, 3, 4), e.g. comparing models with convolutional vs transformer architectures, or with different task objectives. To help explain why the 'mean' makes sense, we have revised our introduction to help clarify the logic of this approach (lines 70-81), and further explained the limitations of the generalization that can be made from these mean-

based statistics in the general discussion (lines 589-592). Indeed, we do think models that are poor predictors of human visual system responses are important, as they give us clues as to which inductive bias might be less brain-like. Weighting by brain-predictivity would confound our independent variables and our outcome measure.

- Second, for the analysis reported in figure 5, we no longer use mean-based approaches, but shift to correlation-based analyses (e.g. with parameter count, ImageNet accuracy, effective dimensionality). You are quite right here that for some of these analyses, like whether effective dimensionality or top-1 accuracy correlate with brain-predictivity, we focus more on the results with trained models, without fully explaining why.
- In the revised manuscript we now clarify the logic of our choice here (lines 400-402 and line 409), which is related to the fact that the effect of training is quite dramatic on both ImageNet accuracy and effective dimensionality. If people want to conclude that effective dimensionality or ImageNet accuracy predicts brain-scores, we reasoned that this relationship should *also* hold among pre-trained models alone. But it does not, even though we have a decent range on the underlying factors (accuracy and dimensionality). So, we think a more accurate account of the data is that neither ED nor Top-1 ImageNet are effective predictors, over and above the benefits of just training the models.
- We hope these revisions help better clarify the logic of our analyses and the scope of the conclusions we can draw from them.

2) The authors are using well motivated single subject based statistics. I think the authors should defend that choice in comparison to statistics based on the group of subjects. There is always ongoing debate on how to best do the analyses for models of fMRI data fairly and how far the results can generalise based on which statistical analyses. Thus, I think a proper defense of the choices made is necessary. The authors do comment on some of these questions already in the supplement, but I believe some discussion in the main text is necessary.

- Thank you for raising this point – we very much discussed this at length as a group at the outset of this project! Here, our choice was ultimately constrained by the particular structure of the NSD fMRI dataset, which comes with a tradeoff between high-density single subject stimulus sampling (e.g. 10000 images for 1 subject) and group-level analyses (e.g. the "Shared1000" images for 4 subjects or the "Special515" for 8 subjects).
- We now (i) directly state this point in the Methods section (in the "Statistical Analyses" subsection, lines 803-811), where we also (ii) discuss the limits on

statistical inference, given our choices, and (iii) acknowledge the debate on single-subject vs large-N group analyses for fMRI design.

Minor points:

As only 4 subjects entered the analyses, presenting those 4 datapoints in some appendix would be a good idea. Even if all the model comparisons were shown this would be a completely readable length.

- Agreed! We have now added a figure to the supplement that shows the scores for all subjects and all models in a single figure (a variation of Figure 5A, see Results lines 360; Section SI.3 and see Supplementary Figure 2).

I think a stronger separation of the different analyses in the plots would be great, i.e. not connecting the boxes with lines above the plots and a little more space between the columns. Perhaps a letter label for the parts. I certainly struggled to keep them apart and in most of them at least, there is space.

- Thank you for this idea! We have adjusted Figures 2, 3, and 4 following your suggestions, with letter labels for the parts of the figures, an attempt at clearer visual separation between the different model set columns, and a more standardized legend schematic. We think the figures are improved, and hope you find them more intuitive, as well.

Reviewer #1 (Remarks to the Author):

I think the revisions have improved the paper, the results are interesting, and I'm happy to recommend publication. But I still think there is some unclarity of what the claims are. Let me just briefly outline where I'm unclear in case the authors want to address this point in any revision.

In a new passage designed to clarify their research agenda, the authors write: "Specifically, we conceptualize each of these DNNs as a different model organism–a unique artificial visual system– with performant, human-relevant visual capacities. As such, each DNN is worthy of study, regardless of whether its properties seem to match the biology or depart from it. We take as our next premise that different DNNs can learn different high-level visual representations, based on their architectures, task objectives, learning rules, and visual "diets". By comparing sets of models that vary only in one of these factors, while holding other factors constant, we can begin to experimentally examine which inductive biases lead to learned representations that are more or less brain-predictive. In our framework, models are not competing to be the best in-silico model of the brain. Instead, we think of them as powerful visual representation learners, with controlled comparisons among them providing empirical traction to study the pressures guiding visual representation formation."

But the authors are not studying DNNs independently of brains (as you would study brains independently of DNNs). Rather, the authors are comparing how well DNNs predict brain activation under different conditions. So, we are not learning anything about "visual representation formation" other than how well the DNNs representations predict brain activation in different conditions. The reason why neuroconnectionism is so popular is the claim is that some DNNs do learn brain-like representations, and as far as I can tell, the authors of this article are claiming to provide some insights into when this is the case. Is that correct?

Relatedly, the authors write that neuroconnectionism "has the potential to unveil the pressures that have shaped the representation we measure in the brain, answering questions about "why" these representations appear as they do (Kanwisher et al., 2023, see also Wood et al., 2020; Vong et al., 2024)." And the authors take their approach to be mostly closely aligned with this research agenda.

But if we want to answer questions about "why" these representations appear as they do" you need to know what features are driving predictions in RSA. And as I noted in my previous review, the design of this research does not provide insight into this. I cited a paper of our work that showed that confounds can drive strong RSAs, such that high RSAs can be obtained between two systems that classify objects based on unrelated visual features. I still think this is a point that merits discussion, and indeed, this is the reason why experiments need to be carried out (as the authors agree in the Discussion). In case the possibility of confounds is addressed in a revision, the reference to this work can be updated as it is now coming out in ICLR workshop of DNN—brain alignment:

Dujmovic, M., Bowers, J., Adolfi, F., & Malhotra, G. INFERRING DNN-BRAIN ALIGNMENT USING REPRESENTATIONAL SIMILARITY ANALYSES CAN BE PROBLEMATIC. In ICLR 2024 Workshop on Representational Alignment. https://openreview.net/pdf?id=dSEwiAENTS

Reviewer #2 (Remarks to the Author):

I appreciate the thought and effort that the authors have put into their responses to the reviews, and find the paper improved. The manuscript presents an impressive body of work, and does a valuable service updating our understanding over earlier efforts that used far smaller stimulus sets and noisier fMRI data. I have no further suggestions or comments.

Reviewer #2 (Remarks on code availability):

Haven't tried running the code but have had a look at the GitHub repository. Seems excellent!

Comes with a Readme and most importantly a really detailed Colab notebook demonstrating the full pipeline for one model. Very helpful resource.

Reviewer #3 (Remarks to the Author):

The authors have adequately addressed my comments and I don’t have further comments.

Reviewer #3 (Remarks on code availability):

Currently the linked repository does contain a readme on how to run the code, but (nearly) none of the code to actually run the analyses, which is actually here:
https://github.com/ColinConwell/DeepDive/tree/main/deepdive

Response to Reviewer #1:

I think the revisions have improved the paper, the results are interesting, and I'm happy to recommend publication. But I still think there is some unclarity of what the claims are. Let me just briefly outline where I'm unclear in case the authors want to address this point in any revision.

- We thank you for accepting our revisions and for all your comprehensive feedback throughout the review process. To prevent further delays to publication, we have largely maintained the manuscript as it was submitted during the last round of revisions. For the sake of the discourse, though, we have also provided point-by-point responses to each of your remaining concerns below.

In a new passage designed to clarify their research agenda, the authors write: "Specifically, we conceptualize each of these DNNs as a different model organism–a unique artificial visual system– with performant, human-relevant visual capacities. As such, each DNN is worthy of study, regardless of whether its properties seem to match the biology or depart from it. We take as our next premise that different DNNs can learn different high-level visual representations, based on their architectures, task objectives, learning rules, and visual "diets". By comparing sets of models that vary only in one of these factors, while holding other factors constant, we can begin to experimentally examine which inductive biases lead to learned representations that are more or less brain-predictive. In our framework, models are not competing to be the best in-silico model of the brain. Instead, we think of them as powerful visual representation learners, with controlled comparisons among them providing empirical traction to study the pressures guiding visual representation formation."

But the authors are not studying DNNs independently of brains (as you would study brains independently of DNNs). Rather, the authors are comparing how well DNNs predict brain activation under different conditions. So, we are not learning anything about "visual representation formation" other than how well the DNNs representations predict brain activation in different conditions. The reason why neuroconnectionism is so popular is the claim is that some DNNs do learn brain-like representations, and as far as I can tell, the authors of this article are claiming to provide some insights into when this is the case. Is that correct?

- The logic you attribute to us here is not quite right. Suppose scientists build a computational model with the explicit aim of *trying to fit brain responses*, and the model fits the responses poorly: this would be a *bad* model. But if a computational DNN model is treated like a 'model organism' with its own representational competencies, then it is still interesting as an existence proof of a solution to gaining those competencies (like object recognition). And so, even if it doesn't fit the brain, it's not a *bad model*. In fact, it might still be a *useful model* for understanding brains, *especially* if a cousin model that is mostly the same but for one different mechanism, is a much better fit to brains. *This* is

how we are aiming to “learn about visual representation formation” by leveraging “how well the DNNs representations predict brain activations”.

- More generally, it is worth noting that the reason that “neuroconnectionism is so popular” could very well differ between research groups that agree to greater and lesser extents with the assumptions of this framework. For us (the authors), we find the primary value of this framework in two of its core premises: (1) that DNNs are powerful learning systems, with competencies we could never model before, and are thus interesting artifacts to study in their own right, and (2) that these models are expressed in a computational language that has coarse algorithmic resemblance to brains (e.g. parallel, distributed, hierarchical, non-linear neuron-like processing, et cetera), and show *emergent* fits to brains (without trying to model brains per se). These premises connect DNNs to levels of analysis that bridge cognitive science and systems neuroscience alike. It also ties them to evo-devo frameworks that emphasize learning mechanisms at least as much if not more than general architectural constraints. Taken together, these premises make clear the value of DNN modeling as a generative research program: Models can easily be built and tested with new hypothesized mechanisms designed to endow them with new capacities, and can subsequently be probed for emergent brain-like properties -- even if, again, they don’t provide “competitive” fits to brains. Many of these properties are highlighted in the “neuroconnectist” research programme, which is why we generally endorse/subscribe to it in our own ways of doing NeuroAI research.

Relatedly, the authors write that neuroconnectionism “has the potential to unveil the pressures that have shaped the representation we measure in the brain, answering questions about "why" these representations appear as they do (Kanwisher et al., 2023, see also Wood et al., 2020; Vong et al., 2024).” And the authors take their approach to be mostly closely aligned with this research agenda.

But if we want to answer questions about "why" these representations appear as they do” you need to know what features are driving predictions in RSA. And as I noted in my previous review, the design of this research does not provide insight into this. I cited a paper of our work that showed that confounds can drive strong RSAs, such that high RSAs can be obtained between two systems that classify objects based on unrelated visual features. I still think this is a point that merits discussion, and indeed, this is the reason why experiments need to be carried out (as the authors agree in the Discussion). In case the possibility of confounds is addressed in a revision, the reference to this work can be updated as it is now coming out in ICLR workshop of DNN—brain alignment:

Dujmovic, M., Bowers, J., Adolfi, F., & Malhotra, G. INFERRING DNN-BRAIN ALIGNMENT USING REPRESENTATIONAL SIMILARITY ANALYSES CAN BE PROBLEMATIC. In ICLR 2024 Workshop on Representational Alignment. https://openreview.net/pdf?id=dSEwiAENTS

- We argue that these measures *are* informative – they just make certain assumptions about what does and does not matter when we call two systems “similar”. What exactly we should and should not assume is indeed a deep philosophical question (e.g. Cao et

al., 2012), but we see this discussion (of how to 'best' link models and brains), both conceptually and analytically, as a conversation that will unfold more over time.

- We entirely agree that a deeper understanding of the features driving brain-alignment metrics should be part-and-parcel of the NeuroAI research agenda moving forward (though there are others who think feature understanding is a red herring). Importantly, the ability we have in task-performant DNNs to access (and *casually* manipulate) not only features, but neural weights, circuits, and populations (far beyond what is currently possible with biological brains) is one of the main reasons we consider them so exciting as a tool for understanding.
- Already in our lab and elsewhere, we have seen these new kinds of analyses being developed and applied to great effect, and in ways that directly lend themselves to the kind of empirical control that has underpinned so much of the behavioral and psychophysical research paradigms that inspire us. In short, the ability to directly follow a through-line from distal ecological constraints (i.e. the "pressures" we have analyzed here") to proximate neural metrics (i.e. features) to behavior (i.e. task-performance) is why we believe DNN models are ideal for asking the "why" questions we want to ask in computational neuroscience.

Response to Reviewer #2:

(Remarks to the Author):

I appreciate the thought and effort that the authors have put into their responses to the reviews, and find the paper improved. The manuscript presents an impressive body of work, and does a valuable service updating our understanding over earlier efforts that used far smaller stimulus sets and noisier fMRI data. I have no further suggestions or comments.

- We thank you for your kind words and all your helpful feedback throughout the review.

(Remarks on code availability):

Haven't tried running the code but have had a look at the GitHub repository. Seems excellent! Comes with a Readme and most importantly a really detailed Colab notebook demonstrating the full pipeline for one model. Very helpful resource.

- We are glad you have found our GitHub to be helpful. Note that we will also now be updating this repo to allow for the reproduction of all results in our analysis.

Response to Reviewer #3:

(Remarks to the Author):

The authors have adequately addressed my comments and I don't have further comments.

- Thank you for your review. We greatly appreciate all the feedback.

(Remarks on code availability)

Currently the linked repository does contain a readme on how to run the code, but (nearly) none of the code to actually run the analyses, which is actually here: https://github.com/ColinConwell/DeepDive/tree/main/deepdive

- We will be updating the *DeepNSD* repo such that it no longer requires interfacing with the *DeepDive* repo in order to reproduce results. Accordingly, all results will be reproducible in the DeepNSD repo.